Best
Available
Copy

AD-785 738

SPEECH DIGITIZATION BY LPC ESTIMATION TECHNIQUES

STANFORD RESEARCH INSTITUTE

PREPARED FOR
ADVANCED RESEARCH PROJECTS AGENCY

JULY 1974

DISTRIBUTED BY:

NTIS
National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 · U.S.A.

AD-785 738

Annual Technical Report – Task 3

Covering the Period 3 October 1972 through 31 March 1974

Form Approved
Budget Bureau No. 22-R0293
July 1974

SPEECH DIGITIZATION BY LPC ESTIMATION TECHNIQUES

By: D. T. MAGILL (Task Leader) and C. K. UN
(415) 326-6200, Ext. 2664

Prepared for:

ADVANCED RESEARCH PROJECTS AGENCY
ARLINGTON, VIRGINIA 22209

CONTRACT DAHC04-72-C-0009
ARPA Order No. 1943
Program Element Code 61101D

SRI Project 1526

The views end conclusions contained in this document ere those of the authors end should not be interpreted as necessarily representing the officiel policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

Approved for public release; distribution unlimited.

Approved by:

ROBERT F. DALY, *Director*
Telecommunications Sciences Center

BONNAR COX, *Executive Director*
Information Science and Engineering Division

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

CONTENTS

Preceding page blank

ILLUSTRATIONS

TABLES

I INTRODUCTION

A. Objective

This study is aimed at the broad goal of the DoD Secure Voice Consortium to develop hardware models of improved narrow-band voice coders. The study is focused on the "pitch and voicing" problem. The objective is to conceive and demonstrate the feasibility of two or more improved strategies to estimate and encode the excitation parameters of human speech. The decoded parameters will be used to excite a time-varying vocal tract "filter" in the synthesizer.

B. Background

1. Difficulty of Pitch Extraction

"Fundamental frequency analysis--or 'pitch extraction'--is a problem nearly as old as speech analysis itself. It is one for which a complete solution remains to be found." Dr. J. L. Flanagan's observation a decade ago remains true today.[1]* Speech analysis-synthesis systems have not gained user acceptance because speech quality and naturalness suffer in such systems. The "machine-like" quality and inability to recognize the talker inhibits user acceptance. Flanagan writes: "The seat of the difficulty is largely the extraction of excitation information--that is, the pitch measurement and the voiced-unvoiced discrimination. The difficult problem of automatic pitch extraction is well known. The device must faithfully indicate the fundamental of the voice over a

---

*References are listed at the end of this report.

frequency range of almost a decade (if male and female voices are to be handled) and over a large range of signal intensity."

2. The Stationary Model

The speech waveform is produced by exciting the linear, time-varying vocal tract filter, $v(t,\tau)$, with an excitation function, $e(t)$, that exhibits a noise-like or periodic impulse character. The observed speech is given by the convolution integral

$$s(t) = \int v(t,\tau)\, e(\tau)\, d\tau \quad . \tag{1}$$

The character of $e(t)$ changes slowly with time, at syllabic rates. Similarly, the vocal tract filter is articulated slowly. Thus, the speech signal may be subdivided into intervals during which both excitation and vocal tract appear stationary. The analysis-synthesis strategy is to segment and analyze successive short-time stationary "frames" into excitation, $e(t)$, and vocal tract, $v(t)$, components for transmission over a reduced capacity channel and subsequent synthesis of the reconstructed speech. Two basic strategies for coding the excitation signal, $\hat{e}(t)$, are:

- Transmit a signal, $\hat{e}(t)$, that contains the natural pitch and voicing structure. The most common example of this strategy would be simply to encode the residual by pulse code modulation (PCM). The adaptive predictive coding (APC) method of Atal and Schroeder is another example.[2]
- Transmit only the coded feature-extracted parameters [pitch frequency and voiced/unvoiced (V/UV) decision] estimated from $\hat{e}(t)$ or directly from $s(t)$. In the synthesizer, $\hat{e}(t)$ is generated from knowledge of the pitch parameters. The most common example of this strategy is the pitch extractor used in the channel vocoder. However, with the linear predictive coding

(LPC) method, pitch can be extracted by performing an autocorrelation analysis on the residual. Transmitting pitch and voicing parameters is efficient, requiring only 150 to 700 bits/s of transmission capacity.

This study seeks to increase the quality of synthesized speech by developing improved concepts and algorithms for estimation and coding a representation of the excitation component of the human speech signal.

C. Scope

This excitation study will concentrate on the feasibility of techniques that process the speech residual. The speech residual will contain primarily excitation information, since the majority of formant information will already have been extracted. The type of formant extractor implemented will be linear predictive.

Two areas of investigation are distinguished by their processing memory: (1) short-term memory and (2) long-term memory. Examples of the former are differential pulse code modulation (DPCM) and adaptive delta modulation (ADM). Examples of the latter are autocorrelation and average magnitude distance function (AMDF) processing.

1. Short-Term Memory

This area of the excitation study considers coding techniques that use one to several residual time samples of memory. This excitation coding takes advantage of short-time redundancy and uses a simple redundancy removal processor. Consistent with this approach are techniques that determine V/UV excitation. In this case, a white noise generator would be used at the synthesizer to produce unvoiced speech.

Due to the restricted memory, development of an effective feature extraction system with short-term memory processing is not possible.

Consequently, modestly high rates are required to encode the residual that is used to generate the excitation function.

2. Long-Term Memory

This area of the excitation study concentrates on the examination of coding techniques that use a time interval of the residual at least 2 ms in duration. These techniques use the relatively long-time correlation of the residual and more complex processors than the short-term memory techniques. Consistent with this area of investigation is the extraction of pitch-pulse location, frequency, and amplitude from the residual autocorrelation function. These parameters, along with voicing decisions, would be quantized for coding and then transmitted to the synthesizer.

D. Outline of the Report

The long-term memory approach results are presented in Section II. Following an introduction, subsections are devoted to (1) delay-lock loop tracking of pitch pulses, (2) deconvolution to obtain the glottal pulse, (3) generalized waveform tracking, and (4) formant-isolation analysis. No complete system simulation is based on the results of the long-term memory approach. Nevertheless, many useful results were obtained and are presented in Section II.

By contrast, a very successful complete system simulation has been performed for short-term memory encoding. Section III is devoted to this system and to several of its modifications. Conclusions based on our research are presented in Section IV.

II LONG-TERM MEMORY APPROACH

A. Introduction

1. Summary of Results

Research in the long-term memory approach has been pursued along two basic paths. With the first approach, the major emphasis is applied to the problem of tracking a deconvolved glottal pulse waveshape extracted from the prediction error or residual signal. With the second approach, the baseband of the input speech is processed so that the harmful effects of formants on pitch tracking are alleviated. We will now consider these two approaches in greater detail.

Initially, pitch extraction could be viewed in the time domain as a problem of time-of-arrival estimated; i.e., precise location in the time domain of pitch pulses provides at least the required information and perhaps more.* The average (short-term) period between pitch pulses is usually sufficient, just as the average (short-term) frequency is usually sufficient. As a result, the feasibility was considered of employing the delay-lock loop to track the glottal waveshape present in the residual.[3] The delay-lock loop, a generalization of the phase-lock loop, is capable of tracking arbitrary waveforms and consequently is appropriate for glottal waveshapes.

---

*A potential advantage of the time-of-arrival approach is that it permits precise placement of pitch pulses on an absolute time scale. Thus, pitch synchronous analysis is possible to perform, if desired.

However, the following problem areas were uncovered. First, it is necessary to deconvolve the effects of the vocal tract to produce the glottal waveshape in the prediction residual, and this may not always be possible. In addition, it may not be possible to find an archetype glottal waveshape that can be used as reference waveform in the delay-lock loop. It is very likely that a waveform can be found that will permit locking. However, as the glottal waveshape changes with pitch, stress, and phoneme, a significant time base lag or shift should result. The lag (or lead) will depend on the true glottal waveshape at that time.

Second, it was noted that the delay-lock loop faced a tremendous acquisition problem. The pitch can vary by almost a decade, e.g., 50 to 500 Hz, posing a serious frequency acquisition problem. Even if it were possible to remove the frequency acquisition problem (by acquisition aids and multiple delay-lock loops), a significant phase acquisition problem exists. The basic problem arises because speech is transient in character, with many starts and stops. Phase-lock and delay-lock loops are basically steady-state tracking systems that are not noted for their good acquisition performance.

Third, the large dynamic range in speech causes some serious implementation problems for the delay-lock loop. For example, for a 50-Hz pitch glottal waveshape, the closed-loop frequency response must be very narrow if it is to avoid the same frequency (100 Hz) present at the correlation multiple output. With this narrow closed-loop response the acquisition time is unacceptably large.

In summary, several serious problems with the delay-lock loop approach were discovered. Most of these problems can be solved. However, it is questionable whether solving these problems represents the most effective overall solution to the overall pitch extraction problem. Most serious of these problems is the requirement that the glottal

waveshape be undistorted in the residual. Consequently, the effort on the delay-lock loop was ceased pending successful generation of a residual with the character of the glottal waveshape.

The next major step was to attempt to deconvolve the effects of the vocal tract transfer function and to produce the desired glottal waveshape. The selected approach was to use a non-Toeplitz LPC analysis during force-free periods of the excitation function. Numerous experiments were performed on synthetic speech. The results were completely successful even when the speech model contained zeros as well as poles. Failure occurred only for the test cases in which a strong level of excitation was maintained.

Based on the above results with synthetic speech, numerous tests were made on real speech. In no case was success achieved. However, in some cases (the most likely to produce the desired results) there was a tendency to generate the desired glottal waveshape. It was later discovered that the reason for the uniform failure* was that the residual needed to be integrated once to account for the differentiation associated with the radiation resistance. If the waveforms had been integrated, the glottal waveshape would have been recovered and the high-frequency noise effects reduced in magnitude. Later tests were performed that demonstrated the glottal waveshape when the residual was applied to a low-pass filter, rather than an integrator. However, the success of this experiment depended critically on the acoustical recording environment.

It was also learned that the glottal waveshape could be recovered in another fashion. Rather than looking for a force-free period,

---

*It was expected that the proposed approach might fail for some or many segments of real speech if the glottal stop did not occur.

one might perform an LPC analysis over several pitch periods so that the excitation could be more nearly modeled as a steady-state, rather than a transient, process. In this case, preemphasis on the speech signal essentially removed the effect of the excitation from the LPC analysis. Thus, the LPC parameters would characterize the vocal tract. The preemphasis, together with the differentiation, produces a 12 dB per octave increase that offsets the average 12 dB per octave decrease associated with the typical excitation source (glottal waveshape). Consequently, the effective excitation is white, and the LPC analysis models only the vocal tract. As a result, the residual when twice integrated produces the desired glottal waveshape. The validity of this approach has been demonstrated on our Interdata 70 simulations.*

Note that either the above-averaging or the force-free approach can produce the desired results on speech recorded under ideal situations. However, if significant phase distortion is present due to room acoustics or electronics, the glottal waveshape may not be recognizable. Thus, pitch extraction techniques based on time-domain waveforms can encounter severe problems. Of particular concern is phase distortion due to acoustic environment, e.g., multipath due to reflections in the room.

The problem with phase distortion suggested yet another concept. Rather than use the time-domain waveform, one might measure the short-term (20 to 50 samples) residual energy. Periodic dips in the magnitude of the energy would be a strong indication that one was in a period of little excitation. Similarly, peaks would be indicative of being in the region of glottal excitation. Even this approach is somewhat

---

*Our simulations were conducted either on the SRI-AI PDP-10 or on the Interdata 70 speech processing facility, whichever was most advantageous for the task.

sensitive to serious phase distortion. If the multipath is sufficiently bad, no force-free periods will exist; in this case no dips in the short-term prediction residual energy will be readily apparent.

Unfortunately, before the short-term prediction error energy approach to pitch extraction could be tested, all resources were redirected to the short-term memory approach. As a result, no further progress has been made with this approach to long-term memory pitch extraction.

The second basic path was to employ formant-isolation techniques on the time-domain pitch extraction problem. One of the problems associated with pitch extraction, or rather pitch-pulse placement, is that destructive interference can occur between the impulse responses from the first and second formants. In this case, placement of the pitch pulses becomes difficult, and nonuniform pulse placement may result even in the presence of constant pitch. This problem is particularly serious when the first and second formants are closely located in frequency.

An approach to this problem is to track the formants and to place narrow-band bandpass filters around each of the two lower formants. Thus, the destructive interference from adjacent formants can be avoided. This concept was tested and tried with some success. It was possible to use a simple pitch-pulse placement algorithm (of the Gold and Rabiner type)[4] on the output of each of the formant-isolation filters. However, some pitch errors did occur and manual intervention was required. It appeared possible to handle many of these problem areas by a properly designed automatic algorithm. Tapes of the resulting quality speech have been demonstrated at several technical review meetings. Obviously, further improvements are required.

The major problems associated with the formant-isolation approach are (1) the loss of time resolution due to the use of narrow-band filters, (2) the possibility of formant errors, (3) the complexity of the procedure, and (4) the fact that the proposed approach does not solve all of the pitch extraction problems. Due to the redirection of resources to the short-term memory approach, the formant-isolation concept has not been pursued further.

2. Summary of the Pitch Extraction Problem

The pitch extraction problem has existed for many years in vocoder research. It has been responsible for unacceptable quality and intelligibility. Although numerous attempts have been made to solve the problem and some progress has been made, reliable pitch extraction still remains a problem.*

The following list presents conditions that makes the problem more difficult than it might appear to the inexperienced researcher:

- Lack of fundamental frequency component
- Phase distortion of the signal
- Background additive electrical noise
- Background acoustic noise
- Multiple simultaneous speakers.

Extraction of pitch from the speech itself may be difficult for one or more of the following reasons:

- Rapid change of formants.
- Rapid change of pitch.

---

*The V/UV decision is included as part of the pitch extraction process.

- A voiced fricative.
- A difficult phoneme, such as /r/, which can possess both turn-on and turn-off excitation components.
- Input speech that is very nearly sinusoidal, causing the residual to be extremely small, since a sinusoidal wave is very predictable.
- Proximity of first and second formants, which may cause destructive interference effects.

As a result of any of the above conditions, the pitch extractor will make errors--possibly in the V/UV decision or in the selected pitch frequency. Typical errors are the choice of a harmonic, or possibly a subharmonic, of the true pitch frequency. Although these errors might occur rather infrequently, the listener is sensitive to these mistakes and to this unacceptable quality that results.

The long-term memory research reported in the following sections is devoted to solving some of the more important problems described above.

3. Fundamental Types of Pitch Extraction Algorithms

Two fundamentally different pitch extractors exist. The first, and most common, is the relative pitch extractor, which is called relative since it determines the pitch periods but not the absolute location of the pitch pulse marks. Thus, only relative (or differential) pitch pulse timing information is conveyed. Autocorrelation, SIFT,[5] and the Gold/Rabiner pitch extractors are examples of the relative approach.

The second is the absolute pitch extractor. It places pitch pulse marks absolutely in time, much as one would do when hand marking pitch pulses. A technique based on peak picking from the time-domain waveform is an example of the absolute approach.

The absolute approach permits synchronous analysis and may also yield improved voice quality by correctly placing the first pitch

pulse in a voiced segment. This could be an advantage for speech with rapid and frequent V/UV/V transitions. Plosives, possibly, can be handled better with absolute pitch extraction. The major disadvantage is that an excessive bit rate is required if the absolute location of each pitch pulse is transmitted, a particularly serious problem for high-pitched speech. Furthermore, a variable (dependent on the pitch) data rate system will result. This is acceptable for asynchronous communication systems, such as packet switching, but it causes considerable problems for conventional synchronous communication systems. In addition, the absolute pitch extraction approach is more sensitive to channel phase distortion than is the relative approach. For example, phase has no effect on the autocorrelation pitch extractors, which respond only to the signal power spectrum.

The relative approach provides synchronous rate pitch information with a low information rate, independent of the pitch of the speech. In most implementations it is not overly sensitive to channel phase distortion. Most relative pitch extractors have inherent smoothing that provides a degree of noise immunity. Disadvantages of the relative approach are that (1) the pitch might be too uniform due to the smoothing, (2) the smoothing window may have problems handling rapid transient phonemes, such as plosives, and (3) pitch synchronous analysis is not possible.

Both relative and absolute approaches are considered in the following sections on pitch extraction. However, it is assumed that only relative information is encoded for transmission over the link since this results in a much lower and synchronous data rate.

B. Delay-Lock Loop Tracking of Pitch Pulses

1. Initial Feasibility Analysis

The feasibility of using a delay-lock loop (DLL) as an automatic pitch tracker has been studied at SRI (see Figure 1).

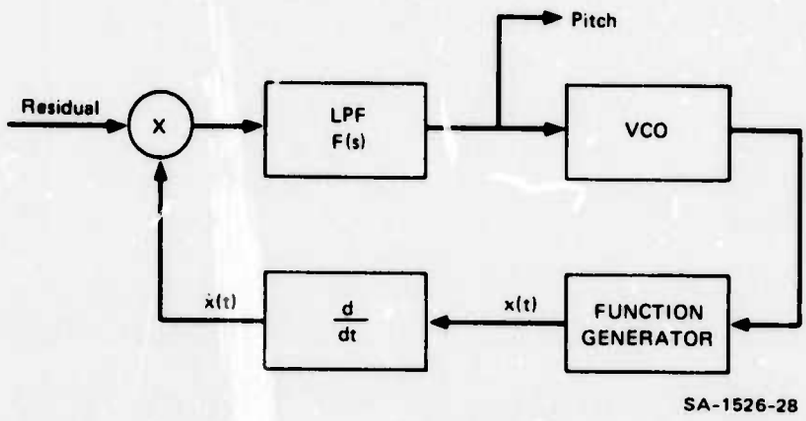


FIGURE 1 DLL PITCH TRACKER

The possible advantages of DLL tracking of pitch pulses are as follows:

- Once acquisition is made, the pitch frequency can be tracked correctly and automatically.
- DLL pitch tracking can be an attractive approach in the presence of background noise.
- If one uses a relative approach to pitch extraction (e.g., autocorrelation) and thus loses the absolute location of the pitch pulses, the pitch pulse placement at the beginning of voiced sounds at the synthesizer could be a problem because no reference signal is available. However, the DLL approach avoids the above difficulty by using the output of the function generator as a reference.

On the other hand, several expected difficulties of using a DLL as a pitch tracking device are as follows:

- The speech or residual waveform is too complex for the DLL to track. To have a proper operation it appears necessary to have a reasonably well shaped unipolar excitation pulse wave extracted from either error signal or speech input.
- Frequency acquisition must be made within several pitch periods. However, according to a preliminary calculation, the acquisition time seems very long, particularly for high pitched speeches.
- The frequency range of speech signal covers many octaves from as low as 50 Hz to over 400 Hz. Yet the maximum frequency acquisition range of a DLL of practical interest is narrow compared with the range of pitch variation.

The discriminator characteristic of a DLL is obtained through the following relationship:

$$\frac{d}{d\tau}\left[R_{ex}(\tau)\right] = \frac{d}{d\tau}\left[\frac{1}{T}\int_{o}^{T} e(t)\, x(t-\tau)\,dt\right]$$

$$= \frac{1}{T}\int_{o}^{T} e(t)\,\frac{dx(t-\tau)}{d\tau}\,dt = -R_{ex}'(\tau) \quad . \qquad (2)$$

Assuming an ideal condition, where we have obtained a glottal pulse wave from speech signals and the function generator generates a similar waveform (see Figure 2), we may plot the discriminator characteristic, $R_{ex}'(\tau)$ (see Figure 3). The region of major interest in the discriminator characteristic is the part having a positive slope, particularly near $\tau = 0$ where the lock will be achieved. Figure 3 shows that the function $R_{ex}'(\tau)$ is approximately linear for small values of τ. One can regard the positive slope region as the essential discriminator that causes the voltage

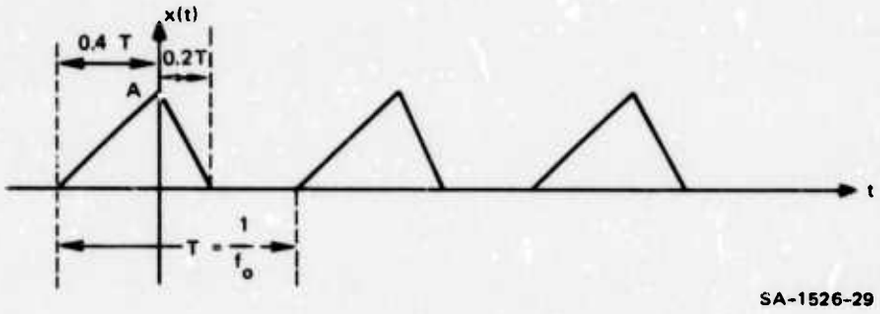


FIGURE 2 IDEALIZED GLOTTAL WAVE

controlled oscillator (VCO) to correct its frequency. Note that the loop may lock onto one of the incorrect positive slope regions, causing error and ambiguity. In this case, the loop will tend to make an optimum estimate, not of the delay between the two signals, but of the delay plus or minus some integral multiple of the pitch period, T.

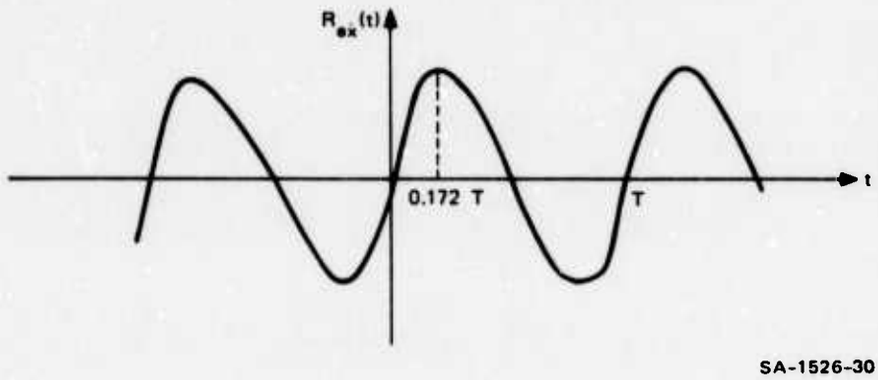


FIGURE 3 DISCRIMINATOR CHARACTERISTICS

With the discriminator characteristic obtained above, a graph of acquisition time versus initial frequency offset, Δf, has been constructed (Figure 4), based on the work of Mengali.[6] The DLL has been assumed to be a second order system with a proportional-plus-integral

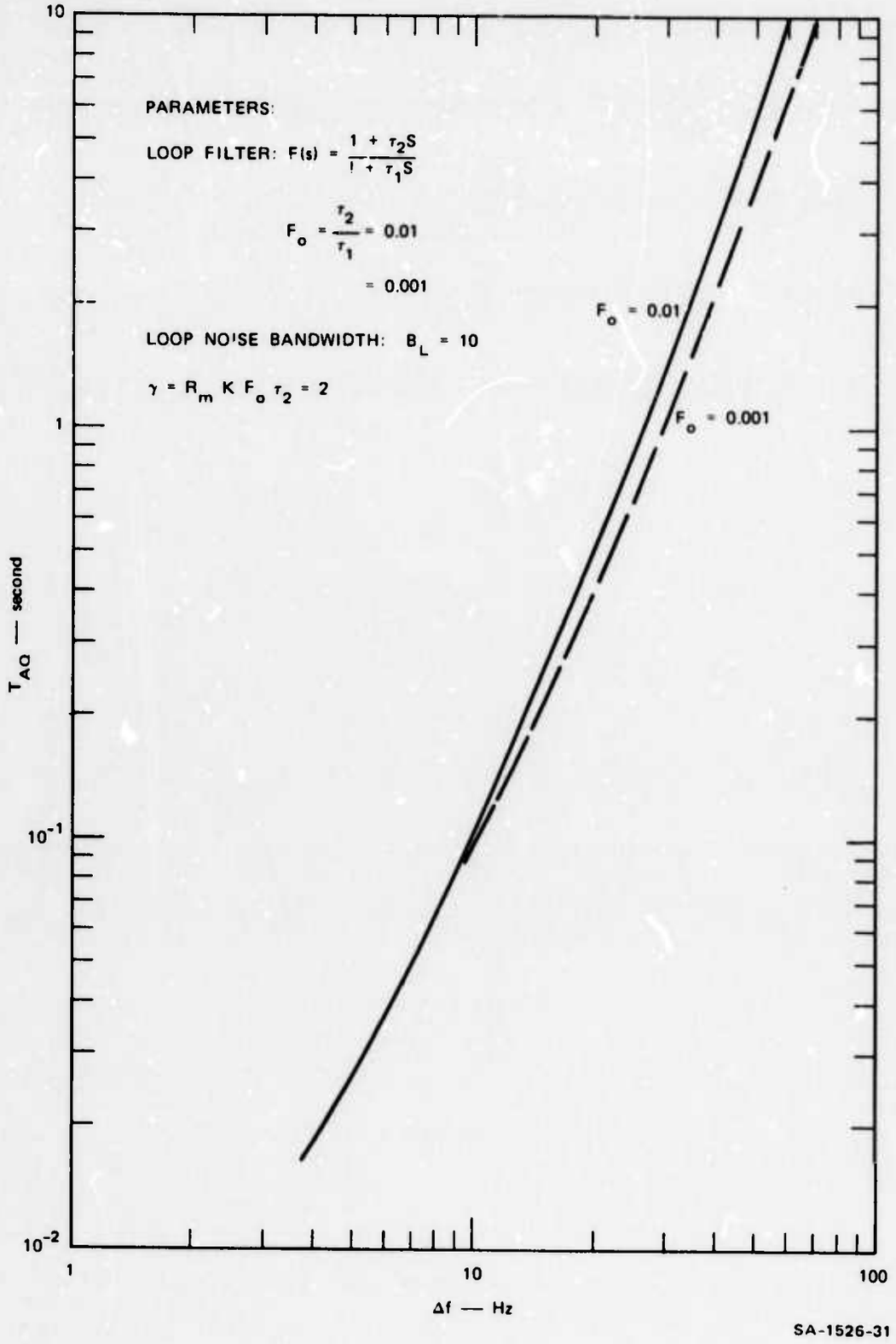


FIGURE 4 ACQUISITION TIME VERSUS INITIAL FREQUENCY DETUNING Δf Hz

loop filter. No noise is assumed to be present in the system. Figure 5 is a graph showing the maximum frequency pull-in range versus the ratio between the ac gain and the dc gain of the loop filter, F_o, using an approximate formula[7]

$$\Delta f_m \simeq \frac{R_m K}{\pi} \sqrt{F_o \left(R_{\dot{e}x}^2 \right)} \qquad (3)$$

where

R_m is the amplitude of the discriminator characteristic

K is the open loop gain

$\left(R_{\dot{e}x}^2 \right)$ is the mean-squared value of $R_{\dot{e}x}(\tau)$.

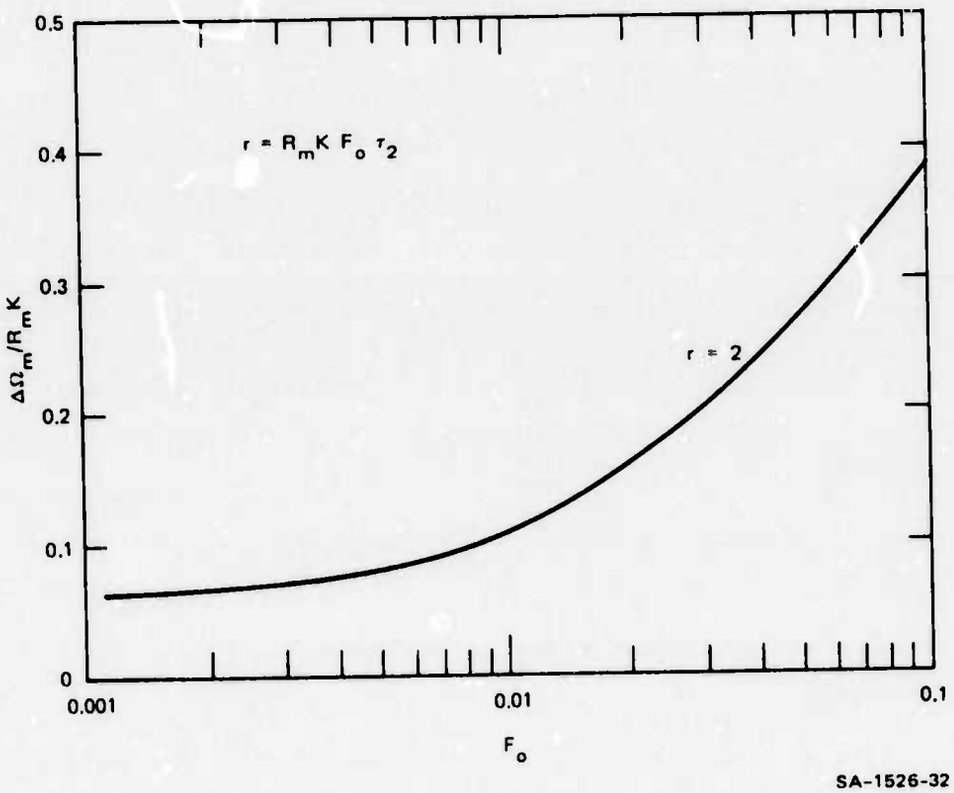


FIGURE 5 FREQUENCY PULL-IN RANGE VERSUS F_o

The frequency acquisition time seems prohibitively long for pitch tracking (Figure 4). For low pitched speech with small initial frequency offset, the problem may not be as bad as expected. However, for high pitched signals, the long acquisition time seems unacceptable even when the frequency detuning is small. One possible method of shortening the time is to use a signal acquisition aid, such as VCO sweeping, or a technique that uses frequency difference measurements.

The VCO sweeping approach cannot be fast enough to solve the problems unless multiple passes (probably more than two) are permitted with the same input data, which requires more and faster computation than is desired. If the frequency difference between the input and the reference signal can be accurately estimated on the first pass, it should be possible on the second pass to avoid the delay-lock loop frequency acquisition problem on the second pass. Nevertheless, a serious delay (or phase for the case of phase-lock loop) acquisition problem exists. For a phase-lock loop, the phase acquisition time is bounded from above by $4/B_L$, where B_L is the equivalent noise closed-loop bandwidth. Depending on waveshape, the delay-lock acquisition time may be considerably more than this bound.

Another disturbing observation (see Figure 5) is that the maximum frequency acquisition range is narrow compared with the fundamental frequency range of speech signals. Consequently, we might need an auxiliary frequency tracker or a parallel frequency tracker with each VCO quiescent frequency set at different values to reduce a large initial frequency detuning. The bank of paralleled delay-lock loops (each tuned to a different frequency) avoids the necessity for a two-pass analysis at the price of greatly increased equipment complexity. Unfortunately, the problem of delay acquisition remains.

In summary, DLL tracking can be an attractive approach to pitch extraction of speech signals, particularly in the presence of background noise. However, for successful implementation of the approach, the following problems must be resolved.

- Extraction of unipolar glottal pulses from voiced speech.
- Acquisition aids for rapid signal acquisition.
- Frequency pull-in range.
- Trade-off between the system complexity and its performance.

The next section presents one approach to alleviating the acquisition problem. Beyond this effort, no further work was performed, pending the successful extraction of glottal pulses from voiced speech.

2. An Acquisition Aid for DLL Pitch Tracking

Due to DLL's long acquisition time and small pull-in range of pitch (note that these two difficulties are contradictory), an acquisition aid or a double pass scheme appears necessary for a real-time DLL operation. One approach is an auxiliary frequency tracker using the autocorrelation method (see Figure 6).

The purpose of an autocorrelator in the DLL tracker is to estimate the fundamental frequency, F_o, of the input signal, X(t), and to feed F_o to VCO so that the initial frequency detuning between the input pitch and the quiescent frequency of VCO can be minimized. Since the role of the autocorrelator is to facilitate a rapid acquisition, its pitch estimation need not be very accurate. However, its accuracy must be within 10 Hz to acquire the signal within one or two pitch periods. Consequently, the proposed autocorrelation will have a relatively less sophisticated decision scheme compared with other algorithms using autocorrelation, e.g., SIFT. The autocorrelator will operate only in the

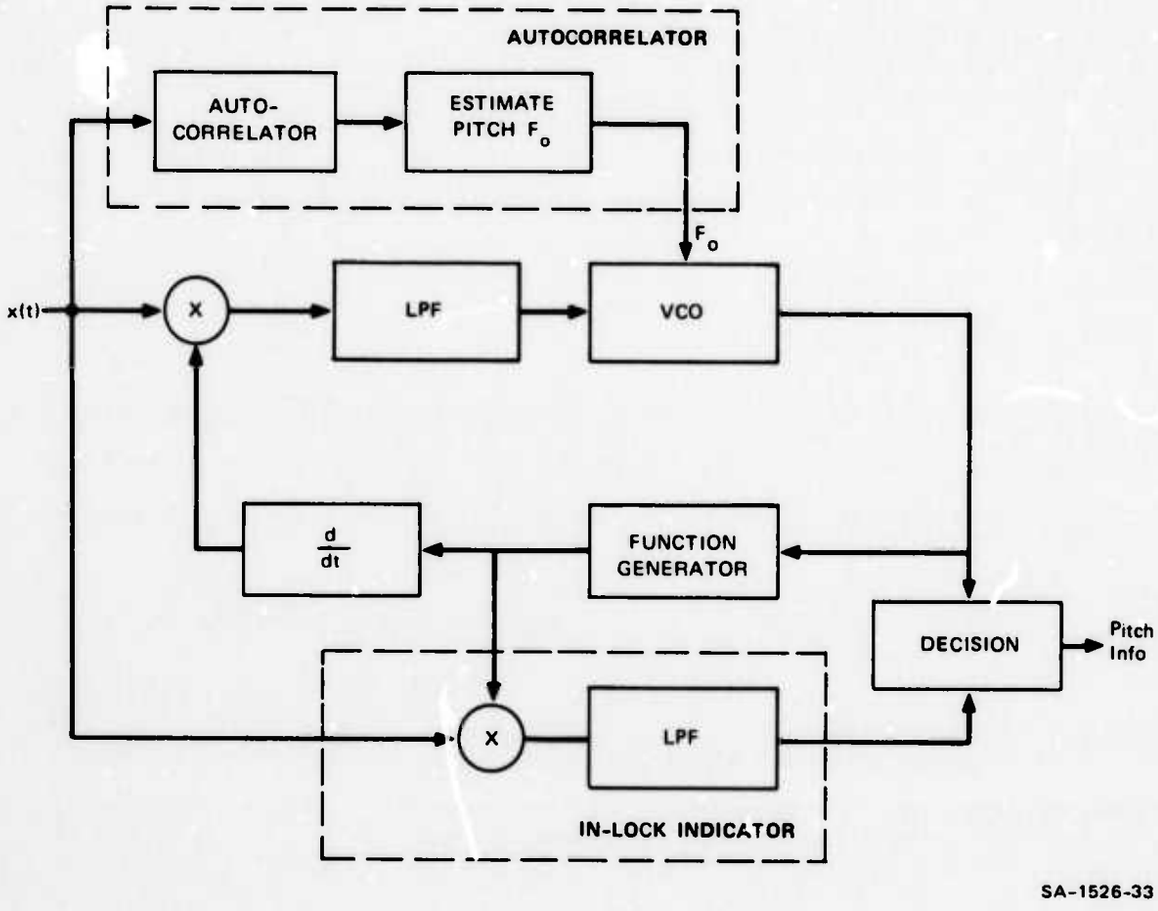


FIGURE 6 DLL PITCH TRACKER WITH AUTOCORRELATION ACQUISITION AID

acquisition mode; i.e., it will be active only when the discriminator characteristic exceeds some threshold value in the case of a large frequency offset. The DLL will thereafter track the voiced input signal, assuming the role of smoothing and fine adjustment for synchronization of phase as well as frequency.

The delay time of the autocorrelator will be set at T = 20 ms (corresponding to the lowest fundamental frequency, 50 Hz) so that the autocorrelation process will be averaged over at least one pitch period of the input signal with the frequency range of 50 to 400 Hz. Because of the delay caused by the autocorrelation process, we need a buffer for the input signal to DLL to achieve synchronization. In addition, we need an in-lock indicator to prevent extraction of erroneous pitch information.

Of course, V/UV decisions are assumed to have been made before the signal arrival at the input of the DLL. Employing the DLL as a V/UV detector does not appear possible. Using an in-lock indicator as a V/UV detector would not be a reliable method because a voiced signal could be out of lock. The V/UV dichotomy can be made separately before the signal arrival by measuring the normalized error energy and comparing it with a threshold or by measuring the zero crossing density of speech input; e.g., when more than two zero crossings/s occur, the speech is classified as unvoiced.*

C. Deconvolution to Obtain the Glottal Pulse

As noted in the previous section, a glottal pulse waveshape in the residual signal must be obtained if the time-of-arrival approach is to be successful. Without this waveshape, design of a delay-lock loop tracking system will be impossible. If it is possible to get the glottal pulse waveshape, then it is also possible to use other, perhaps simpler, pitch extraction routines--simple time-domain peak picking, for example.

---

*Markel claims that the zero crossing density method gives almost 100 percent accuracy for V/UV decision.[8]

The glottal pulse waveshape can be obtained by deconvolving the effects of the vocal tract from the received speech, best accomplished by using an inverse filter, i.e., the filter whose transfer function is the inverse of the vocal tract transfer function. The inverse filter can be best realized as a one-step prediction error filter. The tap gains in the transversal-filter predictor correspond to the LPC parameters.

A difficulty arises in obtaining the proper set of LPC parameters; i.e., unless the correct analysis is performed, these parameters will characterize both the vocal tract transfer function and the glottal excitation spectrum. In this case, the inverse filter will not provide the desired glottal pulse waveshape. The LPC parameters required by the inverse filter are those characterizing only the vocal tract transfer function. A description follows of two LPC analyses that can produce the desired parameters.

1. Force-Free Analysis for Vocal Tract Deconvolution

Linear predictive analysis has frequently been thought of as a statistical analysis procedure for random processes. It is, in fact, nothing more than a regression analysis of an autoregressive random process. However, the speech process is not necessarily a completely random process, i.e., all speech signals cannot be modeled by a white random process driving a linear all-pole filter that shapes the spectrum. In fact, for the large majority of phonemes (e.g., voiced sounds) the speech process is really deterministic. A deterministic forcing function drives a deterministic (but unknown) linear, all-pole filter. Thus, a method is desired to identify the coefficients that describe the deterministic linear filter.

The non-Toeplitz LPC analysis proposed by Atal and Hanauer is a least mean-square error analysis approach.[9] Note that the least

mean-square approach is general and can be applied to either deterministic or random processes. The least mean-square error approach proposed by Atal and Hanauer assumes that the input process is generated by an all-pole filter. If the non-Toeplitz LPC analysis is applied during portions of speech when no forcing function is present, the LPC parameters so derived will perfectly characterize the vocal tract. This assumes, of course, that the vocal tract can be modeled by an all-pole filter. This is often a good approximation.

If the observed speech process is perfectly described by the linear system decay (from 12 initial conditions) of a 12th order all-pole recursive filter, a 12th order non-Toeplitz LPC analysis will characterize the vocal tract perfectly and will produce zero mean-square error in the prediction process. That is, given a past history of 12 samples and one or more new observations, one can predict these new observations perfectly.

The above result can be applied to the problem of pitch extraction. One could attempt to select the analysis interval to correspond to a force-free period and derive the vocal tract characterization perfectly. One could then use these LPC parameters to derive a prediction residual that produces the glottal excitation waveshape. A variety of simple pitch extraction schemes, such as time-domain peak picking, could then be applied. This approach suffers from (1) difficulties in finding the force-free interval, (2) the possible lack of a force-free period for some phonemes and speakers, (3) some phonemes that require zeros as well as poles for perfect prediction, and (4) acoustic and electrical phase shifts that might destroy the existence of force-free periods.

The force-free method of analysis was tried on synthetic speech, using a simple time-share program. A non-Toeplitz LPC analysis was

performed first on speech generated with an all-pole filter. Perfect coefficient estimation and zero prediction error resulted as long as the analyzer included more coefficients than the source model and no excitation was present.

Next, the LPC analysis was tried when the speech model included a few zeros. By shifting the force-free analysis period (in time to avoid the effect of the zeros), it was possible to obtain perfect pole coefficient estimation and zero prediction error. Only a few zeros were used for computational convenience. Obviously, the results could be extended to the required number of coefficients. The true model for voiced speech is the autoregressive moving-average (ARMA) model. The zeros are used to model the excitation waveshape, which can be produced by a series of impulse functions (at the pitch rate) driving a transversal (or finite impulse response) filter. A large number of zeros may be required since the glottal pulse typically occupies 40 percent of the pitch period. For a 10-kHz sampling rate and a 100-Hz pitch signal, 40 zeros are required. A few additional zeros may be required for nasalized phonemes or for other phonemes with acoustic side branches.

Finally, synthetic speech was generated so that a low level (adjustable parameter in the program) excitation was present. Thus, no truly force-free period existed. The LPC analysis was then run at low excitation levels (exact value selectable by operator), and the LPC parameters and the normalized prediction error energy were measured. As expected, the normalized prediction error was no longer zero but increased to a value that was dependent on the level of the constant excitation. The LPC parameters were no longer correct. The magnitude of the error tended to depend on the magnitude of the constant background excitation level.

This final test with synthetic speech was designed to test the sensitivity of the proposed approach to excitation that might always

be present. Typically the glottal pulse occupies from 30 to 70 percent of a pitch period; however, it has been shown that there is great variety in waveshapes. Frequently no true glottal stop exists; this condition was approximated by constant background excitation.

In general, so long as the background level stayed below 10 percent of the peak of the glottal pulse, reasonably good results were possible; i.e., it was possible to use the energy of the prediction residual over a short window (above 20 to 50 samples) as an accurate indicator of the relatively force-free periods. This does not necessarily mean that each LPC parameter is accurately estimated nor does it mean that the residual has a desirable waveshape for time-domain peak picking. In fact, it suggests an extremely attractive possibility for pitch extraction.

This possibility is described in detail in Section II, of this report. The Appendix describes a SUPER BASIC time-share program designed to test the above results; this program was also used to generate the data presented in Section D.

Before leaving the problem of glottal pulse deconvolution, we will consider the following alternative to force-free analysis.

2. Spectral Averaging for Vocal Tract Deconvolution

Rather than attempt to find force-free analysis periods, one can minimize the effect of glottal excitation by performing the proper averaging. First, a large analysis block with several glottal pulses present is required to meaningfully consider the spectrum of the excitation signal. The typical glottal source has a spectral characteristic that falls off at approximately 12 dB per octave. However, the effective value is only 6 dB per octave since the radiation resistance associated with launching the acoustic wave from the lips can be approximated

by a zero. If preemphasized analysis is used, the 6 dB per octave rise with the frequency tends to cancel the effect of the glottal source.* Thus, in theory, it should be possible to reproduce approximately the glottal pulse if an overlapped, Hamming-windowed, pitch-asynchronous non-Toeplitz analysis is performed (with a window size of approximately 20 to 30 ms) on preemphasized speech.

This approach to glottal wave deconvolution was first reported by Allen and Curtis,[10] who successfully demonstrated glottal waveshapes obtained by this method. It is somewhat surprising that this averaging and spectrum cancellation effect works as well as it does. Apparently the LPC parameters that describe the vocal tract can have some error, yet still produce good glottal waveshapes.

A similar system has been employed on our Interdata 70 computer with good results. Figure 7(a) illustrates the glottal wave reconstructed from the residual signal. Note that it does not possess the shape of the true glottal pulse† because (1) no preemphasis was employed and (2) a four-pole, Butterworth low-pass filter with 3-dB cut off at 800 Hz was used instead of a simple integrator. It is interesting that desirable residual waveshapes (for pitch extraction) are obtained in spite of these major differences from the approach proposed by Allen and Curtis.[10]

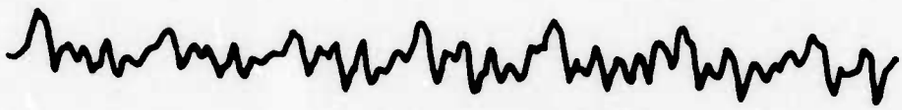
Figure 7(b) illustrates another low-pass filtered residual obtained in the same fashion; the only difference is in the data processed.

---

*In practice, the spectrum of the glottal source does not fall off at exactly 12 dB per octave. Furthermore, the spectrum changes with time depending on phoneme, emotional state, and pitch. However, a reasonably good cancellation is possible in spite of the above problem.

†However, the waveshape is sufficiently simple that a time-domain peak picker will suffice for pitch-pulse placement.

(a) NO PHASE DISTORTION



(b) PHASE DISTORTION

SA-1526-43

FIGURE 7 DECONVOLVED GLOTTAL WAVESHAPES

For this case, the speech was recorded in the far field of a microphone located in a "live" room, i.e., a room with several acoustic reflections. As a result, the low-pass filtered residual is a heavily phase-distorted version of the glottal waveshape. Building a simple time-domain peak picker for locating pitch pulses is not possible.

Thus, if the glottal pulse deconvolution approach to pitch tracking is to be successful, proper precaution is necessary to avoid serious electrical and acoustic phase shifts.

Another possible method of canceling the effect of the glottal source zeros on the LPC analysis is based on the theory of ARMA processes. The theory derived for determining the poles and zeros of an ARMA process can be applied to the present problem. One method of finding the poles is to use the Yule-Walker equations.[11] These equations are essentially identical to the Toeplitz form of the linear predictive equations except

that the correlation coefficients are shifted in time to avoid the effect of the zeros. The equations are

$$R_j \sum_{i=1}^{p} a_i R_{|j-i|} \quad \text{for } i = \ell \text{ to } \ell + p \tag{4}$$

where ℓ is the number of zeros.

Since the required number of zeros is large ($p > 40$), this approach is not particularly attractive. However, this concept might be combined with the previous approach, i.e., preemphasis and analysis based on a large window. In this case, a fewer number of zeros might model the excitation source spectral deviation from a -12 dB per octave characteristic. Thus, for example, one might be able to use the suggested analysis approach of Curtis and Allen, but instead of solving the standard autocorrelation equations, one would solve the Yule-Walker equations for zeros and poles. This combined analysis approach may yield even better LPC parameters for characterizing the vocal tract.

We halted efforts on the above-described hybrid approach, pending consideration of the effects of phase distortion due to acoustic multipath, and so forth, on the residual [see Figure 7(b)].

The next section considers an alternative to deconvolving the effects of the vocal tract and to producing the glottal pulse waveshape as the prediction residual.

D. Generalized Waveform Tracking

As previously noted, a perfectly produced glottal pulse shape in the residual signal may not be possible. For example, phase distortion may prevent the desired waveforms. Our goal here is more modest--merely to produce a signal that permits time-of-arrival estimation.

The concept is best approached from the delay-lock loop point of view. It is possible to construct a delay-lock loop if the received waveform is known. In this section, we present a generalized waveform tracking approach (i.e., an extension of the delay-lock loop) that can operate on arbitrary unknown waveforms. The goal is to find a time-of-arrival discriminant function based on some measure of the input. The only assumption is that the received signal (speech) is produced by an impulse exciting an all-pole recursive filter of order p, or less.

1. Force-Free Theory

This section presents a discriminant technique based on the normalized error energy in the LPC analysis. The LPC analysis used is a special type. The analysis block, N, represents a very few samples, approximately 25. The selected size must be small enough for the block to be phased in time so that only transient decaying waveforms are observed; i.e., a force-free period is processed. The size must be large enough so that adequate data is present to extract 10 to 14 coefficients. A non-Toeplitz LPC analysis was used since it yields correct coefficients, rather than approximations.

Assume that the observed speech waveform is generated by an excitation waveform (whose off-period is at least 25 samples) driving an all-pole p-stage recursive filter where $p < 25$. If the LPC analysis block is phased in time properly, the one-step prediction error energy over this data block should be essentially zero. This is the case since the input waveform is deterministic and the least-squares approach will predict it perfectly. However, if the analysis block includes some excitation, incorrect linear predictive coefficients will be obtained. A much larger error energy will result due to the incorrect coefficients and to the presence of the excitation energy. Thus, it would appear possible to obtain pitch synchronizing information by performing such

an analysis based on each input data sample, i.e., a true sequential analysis.

Figure 8 is a block diagram of the discriminant analyzer. New outputs are obtained each data sample. One output is the discriminant function D(k) given by

$$D(k) = E(k)/R_o(k) \tag{5}$$

where

$$R_o(k) = \sum_{n=0}^{N-1} s^2(k - n) \tag{6}$$

and

$$E(k) = \sum_{n=0}^{N-1} e^2(k - n) \quad . \tag{7}$$

It is inferred that E(k) and e(k - n) are based on $\underline{a}(k)$, i.e., the most recently optimized LPC vector. Consequently, a new LPC analysis must be performed for each data sample.

The discriminant function, or normalized short-term residual power, can be used for pitch extraction by looking for periodic nulls. It is expected that the relatively force-free intervals would occur periodically at the pitch rate. This approach requires a sizable amount of calculation; one LPC analysis per data sample is needed. However, the computational increase is not as great as one would expect; i.e., the amount of computation is not increased by a factor corresponding to the number of data samples per analysis block (approximately 100 to 250). The number of data samples in the autocorrelation evaluation is significantly reduced to approximately 25. Since the autocorrelation evaluation is the largest computational load, this reduction is

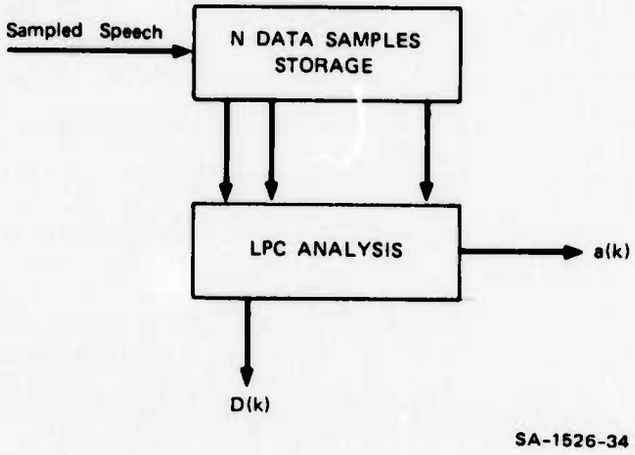


FIGURE 8 BLOCK DIAGRAM OF DISCRIMINANT ANALYZER

significant. Nevertheless, a sizable calculation problem remains. The most realistic method of reducing the calculation is to shift the narrow data window by multiple samples; shifts of 2, 3, 4, or 5 samples are reasonable values to try.

It is worth noting that Sobakin has derived a pitch extractor in a similar fashion.[12] He shows extremely good results for some test phonemes. The major difference is that he uses the determinant of the convariance matrix, rather than the prediction residual energy. There is, of course, a close relation between these two measures. The former is the product of the eigenvalues of the covariance matrix, while the latter is the sum. Thus, low values in one case should lead to low values in the other. Consequently, there is good reason to believe that the approach may work.

The most significant question about the force-free method is how well it will work in the presence of significant acoustic distortion or electrical phase distortion, or both. The prediction model does not

include zeros to represent these effects. As a result, even in a truly force-free interval, a finite prediction residual energy will result. Consequently, the ability to distinguish between forced and unforced intervals will be degraded by the presence of this phase distortion. It is unclear how significant these effects will be in practice. Under most circumstances, they are not expected to create serious difficulties. However, transmission of speech from a remote location over poor quality lines may result in a serious problem.

An alternative approach to the sequential analyzer (or match-filter concept) of Figure 8 is a tracking system. Using this tracking system (once it acquires), one need not compute D(t) for each time sample. Rather, two discriminants--$D_1(t)$ and $D_2(t)$--are used to generate a time-base control voltage. The concept is closely related to delay-lock loop tracking systems where $D_1(t) = D(t + b)$ corresponds to an early channel and $D_2(t) = D(t - b)$ corresponds to a late channel. Figure 9 shows a hypothesized excitation function (its timing is the quantity of interest) and the hypothesized discriminant function, D(t). Note that, during the excitation phase, D(t) is very close to zero. Thus, it should be possible to track the transition (from forced to force-free operation) by requiring $D_2(t)$ to be large and $D_1(t)$ to be approximately zero. Table 1 shows the control voltage that could be applied to a VCO so that the LPC analysis remained in lock at the correct spot.*

Clearly, much more work is required to develop such a tracking system. Questions exist about the control policy and the size of the parameters, such as b, N, and p (the number of poles). Before pursuing this approach, one must establish whether the basic discriminant will function with real speech.

---

*Note that it is not necessary to maintain perfect timing. It is only necessary to keep the analysis block within the force-free phase of the excitation wave.

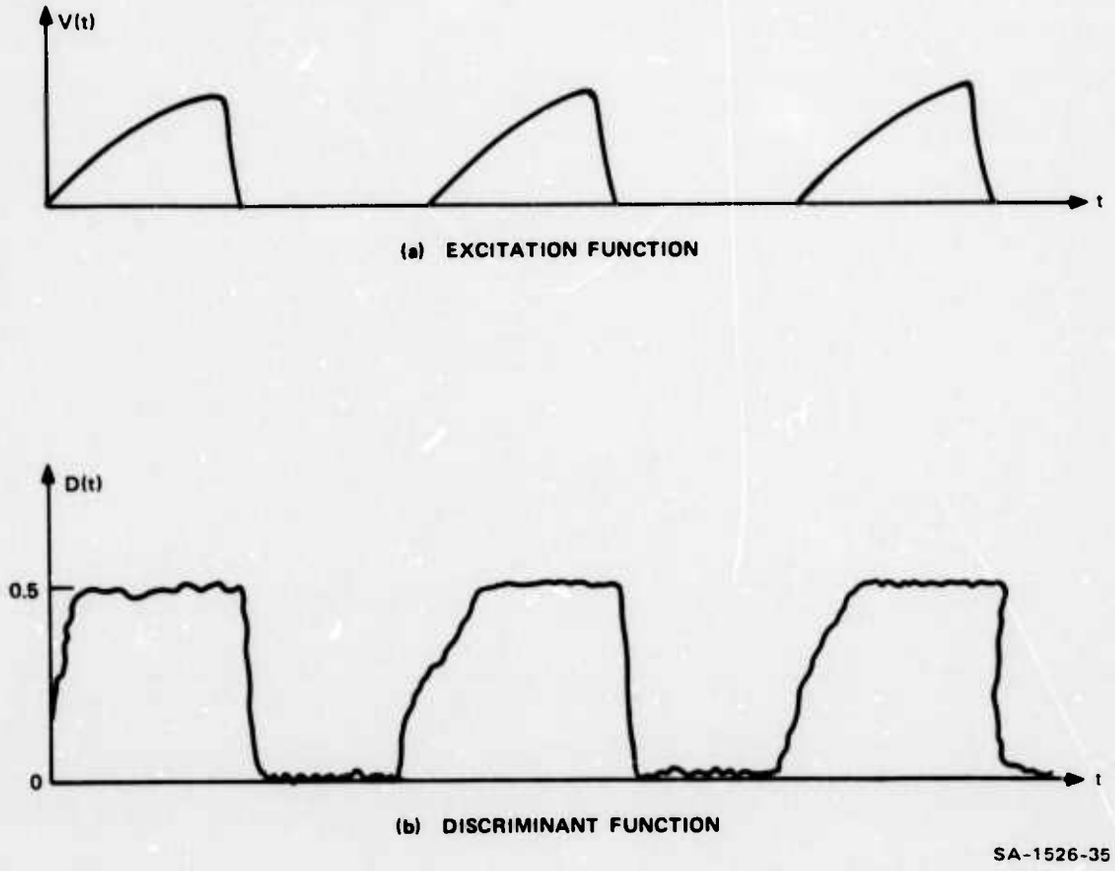


FIGURE 9 HYPOTHESIZED WAVEFORMS FOR THE EXCITATION FUNCTION AND THE DISCRIMINANT FUNCTION ILLUSTRATING THE RELATIVE PHASING

Table 1

CONTROL POLICY

$D_1(t)$	$D_2(t)$	Control Voltage Effect on Time Base
High	High	Speed up
High	Low	Speed up
Low	Low	Slow down
Low	High	Stand still

2. Test Results with Synthetic Speech

Before trying the discriminant concept on real speech, we experimented with synthetic speech so that the desired parameters could be selected and the results checked. Perfect results with synthetic speech having a force-free period should be possible. Our first set of results was obtained with synthetic speech with force-free periods. Later experiments were performed with synthetic speech generated by continuous excitation.

a. Excitation with Force-Free Periods

The synthetic speech was created by driving a linear filter with a very simple excitation waveform. Pulses of unit height were placed to produce a rectangular glottal pulse of the desired width. This crude approximation to the true glottal waveshape was adequate to evaluate the desired effects while being computationally simple.

The following paragraphs report test results obtained under a variety of conditions. Tests were run with speech generated from all-pole and pole-and-zero models. In some cases, the analyzer size was greater than the synthesizer size, in others it was equal, and in the rest it was smaller.

An example of the effect of using an inadequate analyzer size was run to determine if this destroyed the character of the discriminant function. The number of data samples was set equal to 50, while the analysis block size was set at 25. A two-tap analysis was applied to three-tap synthetic speech. The three synthesizer coefficients were 0.25, -0.9, and 0.1. Five pitch pulses were placed at samples 21, 22, 23, 24, and 25. Thus, the first analysis block contained excitation, while the second block was force-free. The discriminant function of the first block was 0.5954 and that of the second block was 0.00585,

showing that a clear distinction can be made on the basis of the discriminant function.

The problem was then rerun with the five pitch pulses relocated at samples 22, 23, 24, 25, and 26. Thus, the second analysis block contained one excitation pulse. The discriminant of the first block was 0.4239 and that of the second block was 0.4687, which showed the discriminant function to be very sensitive to the presence of only one excitation pulse.

As a further check, the problem was run again, but with different synthesizer coefficients. In this case they were selected to be 0.25, -0.8, and 0.15. Otherwise the runs were identical. For the first run the discriminants of the first and second blocks were 0.493 and 0.0314, respectively. For the second run they were 0.429 and 0.708, respectively. Thus, the discriminant continued to be a sensitive indicator of the presence of a forcing function, or excitation pulse.

Test cases were run with synthetic speech containing zeros obtained by passing the output of an all-pole recursive filter through a transversal filter. The number of taps and their values are controllable as program inputs. For all cases run, two taps with weightings of one and two were used.

For all runs, the number of data samples was 50 and the block size was 25; the four synthesizer coefficients were 0.25, -0.9, 0 and 0. (Note that this is really just a two-tap synthesizer.) The analyzer size was two (unless otherwise noted).

For the first run, five pitch pulses were placed at location 21, 22, 23, 24, and 25. The discriminants of the first and second blocks were 0.359 and 0.352, respectively, thus proving that the discriminant is not a reliable indicator in this situation. However, the

next run tested an idea related to the effect of zeros that showed the discriminant could still be used.

In the second run, the five pitch pulses were shifted by one sample to locations 20, 21, 22, 23, 24. In this case, the discriminants of the first and second blocks were 0.384 and +1.8 E-10, respectively.

In the third run, the analyzer was increased to size four while the pitch pulses were shifted back to locations 21, 22, 23, 24, and 25. The discriminants of the first and second blocks were 0.476 and 0.8770, respectively. Thus, increasing the analyzer size did not help.

In the fourth run, the analyzer size remained at four while the pitch pulses were shifted by one to locations 20, 21, 22, 23, and 24. The discriminants of the first and second blocks were 0.171 and 1.7 E-10, respectively. Thus, good discrimination remained even when the order of the analyzer exceeded that of the synthetic speech.

One approach to designing a robust analysis system is to use an underpowered analyzer. The hypothesized advantage of this approach can be best described as follows. If the analyzer is overpowered by the input speech, it will tend to use its extra poles to model the excitation wave in the input speech. In this case, the error energy discriminant might not be large enough to permit separation of forced and forced-free periods.* To avoid the problem, one may wish to use an underpowered analyzer. The basic question is how sensitive our discriminant function is to matching the analyzer size to the dimensionality of the input signal.

---

*On reflection, this does not appear to be a serious threat. In spite of the motivation, the experiment described is of interest.

For this series of three runs, the number of data samples was 50 and the analysis block size was 25. The synthesized signal contained two complex pole pairs. The four coefficients were 1.2, -1.66, 0.972, and -0.689. The analyzer size was two.

In the first run, five pitch pulses were located at samples 21, 22, 23, 24, and 25. The discriminants in the first block and second block were 0.213 and 0.124, respectively. Thus, some separation between forced and force-free periods was maintained even though the analyzer was underpowered.

In the second run, the five pitch pulses were shifted by one sample to locations 22, 23, 24, 25, and 26. Thus, no force-free analysis blocks occurred. The discriminants in the first and second blocks were 0.146 and 0.186, respectively. Note how sensitive the discriminant was to the presence of a single excitation pulse.

A third run was performed with the pitch or excitation pulses shifted to 23, 24, 25, 26, and 27; i.e., two excitation pulses existed in the second analysis block. The discriminants in the first and second blocks were 0.102 and 0.180, respectively. In spite of three excitation pulses in the first analysis block, it produced a lower discriminant than the discriminant for the second analysis block (force-free) in the first run. This discouraging result indicated that use of an underpowered analysis is not desirable; it is preferable to overpower the analysis if the correct value is not known.

Sensitivity to oversizing the analyzer was tested by running a computer simulation with a two coefficient (0.25 and -0.90) synthesizer and an analyzer of dimension four. Fifty data samples were divided into two analysis blocks of dimension 25 each. Five pitch pulses were placed at locations 21, 22, 23, 24, and 25. The discriminant for the first analysis block was 0.3203; for the unforced second analysis block,

the discriminant was 1.7 E-10. Thus, having an overpowered analyzer did not seem to harm the discriminant functions' ability to identify force-free intervals.

The discriminant approach to pitch-pulse tracking was tested on synthetic speech generated by a two-tap synthesizer with coefficient values of 0.25 and -0.95. For all runs the number of data samples was 50 and the analysis block size was 25. The analyzer was dimension two, and five contiguous pitch or excitation pulses were used for all runs, except runs five and six. The runs differed in the location of the pitch pulses.

The results of these computer simulations are presented in Table 2. The force-free analysis blocks can be identified by a discriminant of essentially zero. Any analysis block containing an excitation pulse has a discriminant greater than or equal to 0.1485.

Table 2

SIMULATION RESULTS

Run Number	Pitch Pulse Locations	First Block Discriminant	Second Block Discriminant
1	5, 6, 7, 8, 9	0.590	1.67 E-10
2	21, 22, 23, 24, 25	0.639	1.67 E-10
3	22, 23, 24, 25, 26	0.453	0.295
4	23, 24, 25, 26, 27	0.371	0.462
5*	1, 24, 25, 26, 27, 28	0.305	0.392
6*	6, 29, 30, 31, 32, 33	0.1485	0.459

*Run with six pitch pulses, one of which is separated from the main group of five pulses.

b. Continuous Excitation

The SUPER BASIC time-share program was modified to permit the excitation to have a fixed but selectable level during the "off" period. The nominal excitation level during the "on" period was set to unity. The off level could be chosen to be any less than unity. Since the off level was represented by a constant, it also affected the on level. Thus, the true on level was always greater than one, being the sum of the off level plus one.

Experiments were performed for three synthesizing filters. For Case 1, the synthesizing filter was described by the parameters $a_1 = 0.3$ and $a_2 = 0.9$. For Case 2, $a_1 = 0.3$ and $a_2 = 0.97$. Both of these tests simulated single-formant speech. Case 3 simulated two-formant speech; the selected values were $a_1 = 1.2$, $a_2 = 1.66$, $a_3 = 0.972$, and $a_4 = 0.689$. In all cases the major excitation was present for a portion of analysis block one, while block two contained only the off-level excitation. Details of the program that generated these results are present in the Appendix.

Table 3 presents the simulation results obtained for Case 1 for two levels of off-period excitation. For a constant 0.01 off-level excitation, good results were obtained for both analysis blocks; i.e., the estimated LPC parameters were very close and the discriminant function during the force-free period was much lower than for the period when the major excitation was present. However, for the case of a constant 0.10 off-level excitation, the results were not nearly so good. Surprisingly, the LPC parameters were farther off in the relatively force-free period than during the period of major excitation. However, the discriminant function was lower during the force-free period, as predicted. Unfortunately, the difference between the two discriminants was not nearly so large as for the preceding case of weaker off-period excitation. Nevertheless, a distinction did appear possible.

Table 3

CASE 1--SIMULATION RESULTS

Off-Period Excitation Level	Analysis Block	a_1	a_2	Discriminant
0.01	1	0.3016	-0.8994	0.2317
	2	0.3009	-0.8943	1.7 E-03
0.10	1	0.3295	-0.8811	0.2699
	2	0.3691	-0.8029	0.1399

Table 4 presents the test results for Case 2. Here simulations were run for off-period excitation levels of 1 E-04, 1 E-03, 1 E-02, and 1 E-01, respectively. Table 4 shows that for all cases, good results were obtained by using the discriminant function to isolate force-free periods.

Table 4

CASE 2--SIMULATION RESULTS

Off-Period Excitation Level	Analysis Block	a_1	a_2	Discriminant
0.0001	1	0.3142	-0.9605	0.2317
	2	0.3329	-0.9527	2.8 E-02
0.001	1	0.3000	-0.9700	0.202
	2	0.3002	-0.9699	3.28 E-06
0.01	1	0.3005	-0.9699	0.2044
	2	0.3024	-0.96933	3.25 E-04
0.1	1	0.3142	-0.9605	0.2317
	2	0.3329	-0.9527	2.8 E-02

The results for Case 3 are presented in Table 5 for off-period excitation levels of 1 E-05, 1 E-03, 1 E-02, and 1 E-01. Good separation of off and on periods was provided by the discriminant function for all cases except for the highest level of off-period excitation. Separation is still possible, but the difference between the two values is not as high as desired. It is not clear that separation is possible for all possible speech waveforms. However, success was achieved for all cases.

Table 5

CASE 3--SIMULATION RESULTS

Off-Period Excitation Level	Analysis Block	a_1	a_2	a_3	a_4	Discriminant
0.00001	1	1.2000	-1.6599	0.97200	-0.6889	0.1659
	2	1.2000	-1.6600	0.9720	-0.6890	1.24 E-08
0.001	1	1.2002	-1.6598	0.9720	-0.6885	0.166
	2	1.2028	-1.6652	0.9762	-0.6918	1.25 E-04
0.01	1	1.2035	-1.6579	0.9725	-0.6844	0.1677
	2	1.2627	-1.7273	1.0286	-0.7011	1.17 E-02
0.1	1	1.2716	-1.6362	0.9912	-0.6153	0.196
	2	1.95	-2.0791	1.3497	-0.4218	0.1063

Figure 10 plots the discriminant function (i.e., the normalized residual energy) during the off period as a function of the off-period excitation level. On the basis of the test cases, if the real speech processed has off periods where the excitation is less than ten percent of the peak glottal pulse, it should be possible to recognize these periods from the discriminant function.

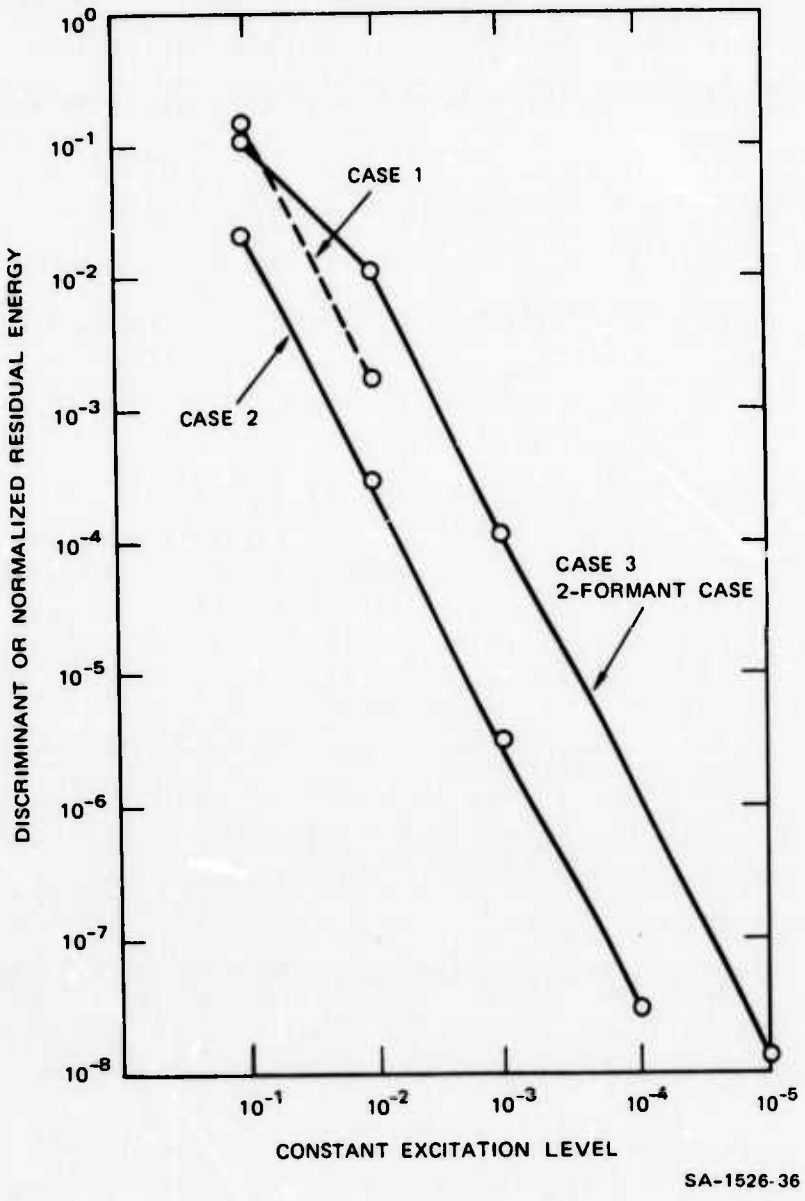


FIGURE 10 DISCRIMINANT DURING "OFF" PERIOD AS A FUNCTION OF THE CONSTANT EXCITATION LEVEL

c. Summary of Test Results

Extremely successful results were obtained with synthetic speech. Consequently, tests were run on real speech; unfortunately, very poor results were obtained. Difficulties with some speech segments

had been expected. However, experiments could not obtain the desired result (a successfully deconvolved glottal pulse) for any segments of real speech.* For a few segments a tendency to produce the glottal pulse appeared, but it was heavily masked by spurious high-frequency oscillations.

As a result of these discouraging results, the force-free discriminant function approach was temporarily dropped. Later we discovered the reason for the complete failure was that the residual required integration to cancel the effects of the radiation resistance. Without this integration, the high-frequency noise effects caused by imperfections in the analysis and in the modeling are not sufficiently suppressed, and the glottal pulse shape is distorted due to the differentiation associated with launching the acoustic wave from the lips. Fortunately, the integration requirement was discovered before subsequent testing of the spectral-averaging method; thus, this approach did not encounter the problem of masking by spurious high-frequency oscillations.

In principle, the spectral-averaging approach is less sensitive to the character and the details of the glottal source than is the force-free method. Consequently, our research efforts were concentrated on the former approach and are reported in the following paragraphs.

3. Discriminant Approach Based on Spectral Averaging

The force-free method of discriminant analysis could possibly be made to function for many speakers and many phonemes, but it is

---

*These tests consisted of observing the residual waveform rather than calculating the discriminant function.

doubtful how well it would work for all speech. Furthermore, it has the disadvantage of requiring a large amount of computation.

An attractive alternative is suggested by the spectral-averaging approach proposed and tested by Curtis and Allen.[10] Here we perform a Toeplitz, Hamming-windowed LPC analysis over a large window (approximately 25 ms) on preemphasized speech. The LPC parameters so derived are used in the inverse filter to form a residual signal. The discriminant function is then evaluated by measuring the moving average of the normalized (with respect to the signal power) residual energy over a short block of 20 to 30 samples. This procedure has several advantages. First, it requires far fewer calculations than does the force-free period method. Second, it should be less sensitive to phase distortion effects. Third, it should be less sensitive to the presence of constant excitation due to incomplete glottal stops.

Note that the spectral-averaging discriminant function approach is fairly straightforward. A special LPC analysis is performed and the residual created. The integrated residual is treated by a simple nonlinearity, the square-law operation, to emphasize the peaks and then is averaged over a modest number of samples. This operation is not much different from some of the ideas presented by Atal and Hanauer[2] for time-domain pitch extraction. They suggest using a cubic nonlinearity to emphasize peaks, but they do no averaging. We expect that averaging would improve performance significantly.

We terminated this work to concentrate on the short-term memory encoding approach. Consequently, the discriminant approach was not pursued with real speech. However, for further studies, the spectral-averaging approach appears to be clearly superior to the force-free approach.

E. Formant-Isolation Analysis

In this section we describe an approach to time-domain pitch extraction that attempts to avoid one major problem in pitch-pulse placement. Hypothetically, a problem exists when the first and second formants of speech are close in frequency. In this case, depending on the phase relations between the two formants, destructive interference can result. The result is that the signal peaks may wander even though the pitch is constant. In short, for these phonemes it is difficult even for a human to correctly place pitch pulses.

The intent of the formant-isolation approach is to separate the individual formants so that pitch extraction can be performed on each of the formant frequency pass bands separately, without interference from signal energy at the other formant frequencies. In this case, accurate placement of pitch pulses should be possible. Isolation is achieved using bandpass filters centered at each of the first two formants, which requires that the first two formants be extracted by some method. The proposed approach uses peaking-picking routines based on the power spectrum envelope derived from an LPC analysis. Figure 11 is a block diagram of the formant-isolation pitch extractor.

The bandwidth of the formant-isolation filters should be narrow enough to avoid the other formant but should not be too narrow, if time resolution problems are to be avoided. In some cases, varying the filter bandwidth between two values to accommodate very close formants may be desirable.

A major goal of the formant-isolation approach is to simplify waveforms so that simple time-domain pitch extractors can be applied on each formant. While these simple pitch extractors may not function perfectly, it is hoped that a decision circuit operating on both extractors may correct the errors made by any one channel. Thus, the system is similar

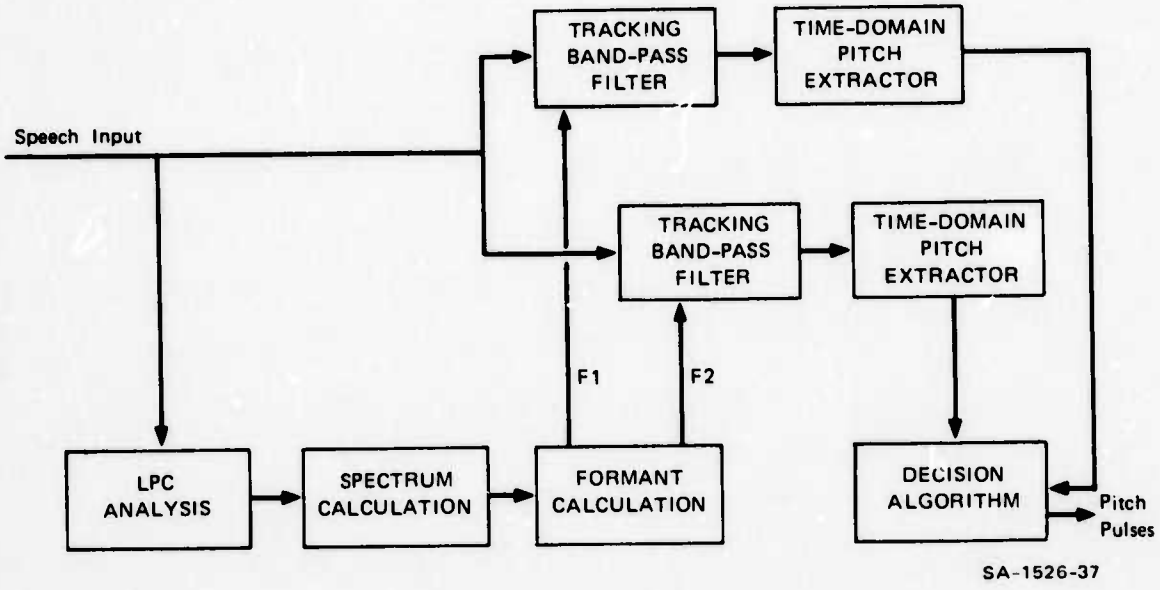


FIGURE 11 BLOCK DIAGRAM OF THE FORMANT-ISOLATION PITCH EXTRACTOR

to the parallel processing concept of Gold and Rabiner.[4] In fact, we have adopted many of their time-domain pitch extraction concepts.

1. Formant-Tracking Filters

The primary motive for filtering the acoustical signal before extracting pitch is to provide a "cleaner" signal while preserving pitch information. We hypothesize that proper filtering will improve pitch detection in the presence of noise and when amplitude and phase distortion

occurs, e.g., in phone circuits. However, the speech signals treated here have high signal-to-noise ratio (SNR)--greater than 40 dB--and minimal distortion. Thus, we will not consider SNR improvement and distortion at this point.

Speech signals transmitted over phone circuits with bandwidths of 300 to 3000 Hz still retain perceptual pitch. Fant's model for non-nasalized vowels indicates the signal energy in this frequency range can be attributed to three or four poles or formants.[13] The time history of these primary formants is called the F pattern, where the formant pole traces are denoted F1, F2, F3, and F4. In the frequency domain, harmonics of the pitch frequency close to these pole locations are emphasized. The time-domain interpretation (subsequently discussed) follows from Fant's cascade, four-stage vocal tract filter, which can be transformed into a parallel filter by a partial-fraction expansion. Each stage then corresponds to a bandpass filter with the formant frequency as center frequency and the glottal pulse as excitation.* Isolation of each branch's output should give a cleaner waveform for pitch extraction even when no fundamental is present; i.e., interference from the other formants will be eliminated by the formant isolation.

a. Filter Characteristics

A bandpass filter centered at the formant frequency with linear phase (to reduce distortion due to changing center frequency) and with steep skirts (to reject other formants) will emphasize the time response from one branch or channel. A Lerner filter design was used to specify a digital recursive filter with the desired properties

*It is assumed that each formant is excited simultaneously with the others.

of linear phase and steep skirts.[14] For a three-stage filter, one pole pair is placed at the desired center frequency, f_c, with the other two pole pairs placed $\pm 2a$ from f_c, respectively. All pole pairs have real parts equal to b. The residue of the center pole pair is b and -b/2 for the two outside pole pairs. Thus, the center frequency, f_c, and bandwidth, $b_w = 2a$, completely specify the filter parameters when the ratio b/a is set. With this ratio the theoretical equation for in-band attenuation and phase is

$$|H(\omega)| \approx \pi b\eta/a \ (1 + \eta^2 \cos \pi\omega/a + \ldots)$$

$$\angle H(\omega) \approx -\omega\pi/(2a) - \eta^2 \sin \pi\omega/a + \ldots \qquad (8)$$

where

$$\eta = \exp(-\pi b/2a) \quad .$$

Thus $H(\omega)$ approximates a constant magnitude with a periodic ripple of relative magnitude, η^2, and a constant time delay of $\pi/2a$ with a maximum periodic error of η^2. For b/a = 1.5 the theoretical amplitude and phase errors are ±0.9 percent and ±0.6 degrees. The filter skirts give 40-dB rejection at $f_c \pm 6a$ Hz.

The importance of linear phase filters can be experimentally demonstrated. The center frequency (and possibly the bandwidth) of each formant isolation filter must be changed periodically. The changes cause transient effects that can produce errors in the simple time-domain pitch extractors. Linear phase filters appear to minimize these transient effects (see Figure 12).

The bottom trace of Figure 12(a) shows the output of a three-stage, 500-Hz bandwidth, bandpass Butterworth filter. The sharp discontinuity occurs at the point when f_c changed from 546 to 351 Hz.

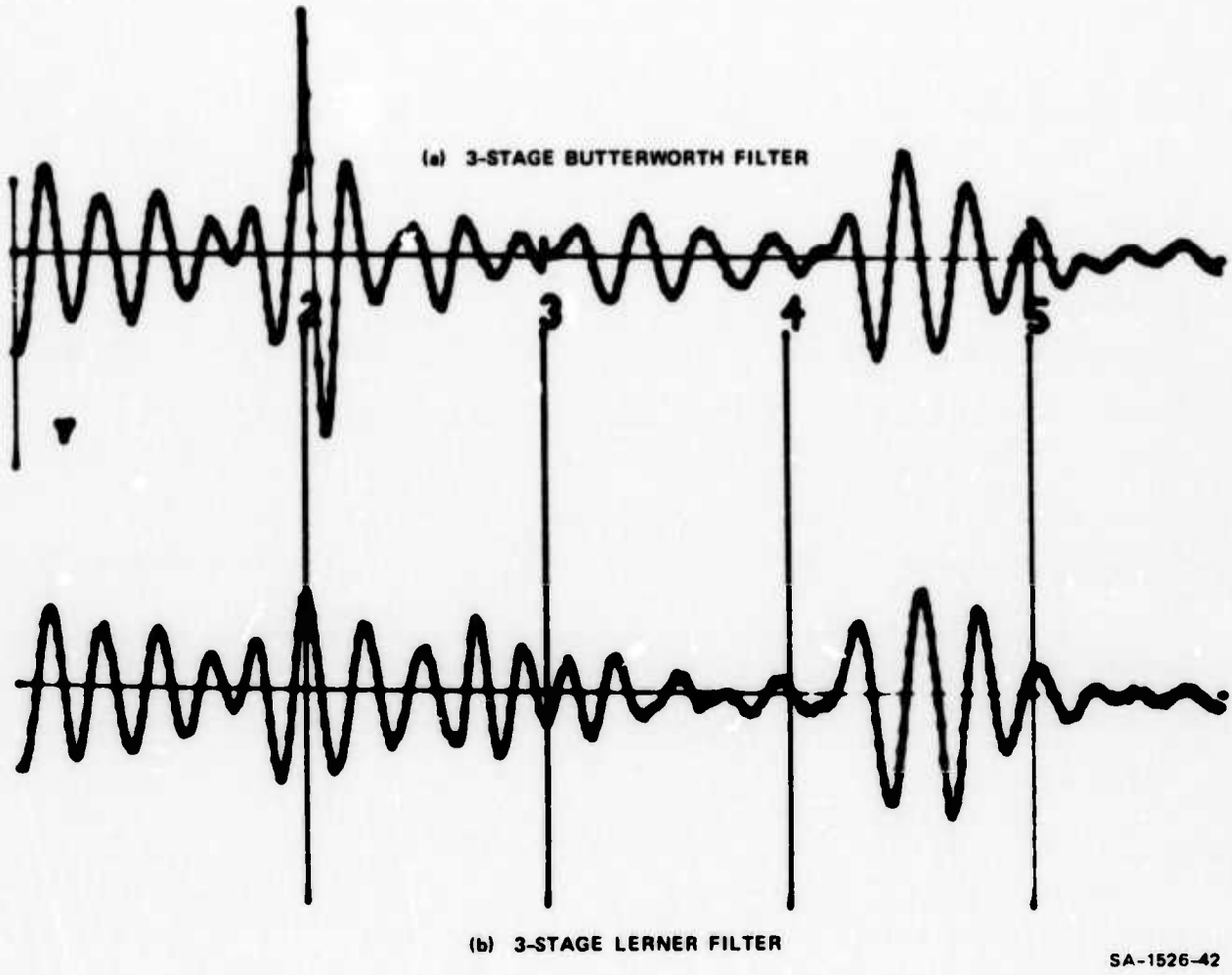


FIGURE 12 OUTPUT SIGNALS FROM FORMANT-TRACKING FILTERS

The bottom trace of Figure 12(b) shows the output of a three-stage Lerner filter of the same bandwidth processing the same speech segment. Note the absence of a pronounced transient. The superior performance of the Lerner filter is due to its superior group delay-versus-frequency characteristic.

b. Formant Tracking

The most critical problem for tracking filters operating on speech signals is estimating the filter center frequency. Formants can be easy to estimate and track during nonnasalized vowels, where Fant's F pattern is a reasonably complete description of the signal characteristics. However, during nasals, liquids, nasalized vowels, and a few other voiced speech segments, additional poles and possibly zeros must be added to the model for a complete description. The F pattern is still present and varies continuously, but the zeros may mask one or more F-pattern poles so that detection is hampered; i.e. formant tracking in general is a most difficult problem. For our purposes here, however, completely solving the formant tracking problem is not necessary. We do not need correct labeling or ordering of the formants; we need only the frequency regions of high energy.* Another simplifying factor is that we do not need to estimate all of the formant frequencies; the first and possibly the second are sufficient. In spite of these simplifications, a difficult problem remains. Simple peak picking on the short-term spectrum is not enough.

Nevertheless, we are interested in a reduced-complexity, formant-tracking algorithm, which is required if the formant-isolation

*Correct ordering can be difficult when formants merge.

approach is to be practical. One algorithm can be stated very simply for the lowest frequency tracking filter; i.e., track the pole corresponding to the first energy peak in the short-term spectrum. In addition, we impose the constraint that the center frequency must vary continuously. Presently, a Newton-Raphson polynominal root-finding technique is used for the polynominal formed from the LPC coefficients for one analysis epoch. However, other algorithms could be used. Note that the frequency estimates need not be exact; this may permit algorithm simplification. The estimates are checked over three successive epochs for continuity. If the middle estimate is greater by a set of threshold value than the previous sample and if the previous and next sample are within a threshold value of each other, then the middle sample is replaced by the average of the previous and next sample. This is the only smoothing required to give good tracking for F1. This algorithm is slightly different from more conventional techniques (see McCandless)[15] in that estimates are not averaged to smooth the transitions unless a sample is skipped.

Generating an estimate for the next energy peak (which may be F2 in some cases) is much more difficult. The second formant changes much more rapidly over a wide frequency range. In addition, the nasal formant is often mistaken for the second formant on the basis of energy peaks. Although one can clearly identify a set of poles in the appropriate region, finding a smooth frequency track is not always possible. Successful smoothing algorithms generally require more samples (than the preceding and succeeding) to adequately smooth the frequency estimates and to select the proper poles.

An example of a non-real-time smoothing algorithm that tracks F2 well enough for our purposes (i.e., it finds a smooth frequency estimate) is the following:

- Find the tie points in the second (ordered in frequency) pole trace. A tie point is a region (say about 30 to 50 ms) where the pole estimates are well behaved, e.g., the region of transition out of a voiced interval.
- Fit a V pattern (two straight lines joined at one end) to the second pole estimate between two tie points and compute the variance about this fit.
- Apply smoothing if the variance is above the threshold; otherwise proceed to the next two tie points.
- Look at second and third pole estimates (i.e., use the smoothing rule) and select one that is within a threshold of the V pattern. If neither are, use the value of V pattern.
- Use a measurement of the percentage of the time pattern value to tell how good the fit is. For poor fits, find a new tie point between the two being used and repeat the process. Good results were obtained with this algorithm for difficult speech segments, e.g., erratic pitch pulse segments, nasals, and liquids. However, it must be emphasized that this is a complex smoothing algorithm that was performed in non-real-time.

In conclusion, tracking of the first formant is not difficult and essentially no errors occur; tracking the second formant is much more difficult and requires a complex algorithm. However, the formant-isolation pitch tracker may be capable of operating in the presence of a formant-location error. One method of improving the system performance is to use a fixed low-pass filter for a baseband pitch extractor and a tracking bandpass filter for the first-formant pitch extractor. Thus, difficulties in tracking the second formant are avoided.

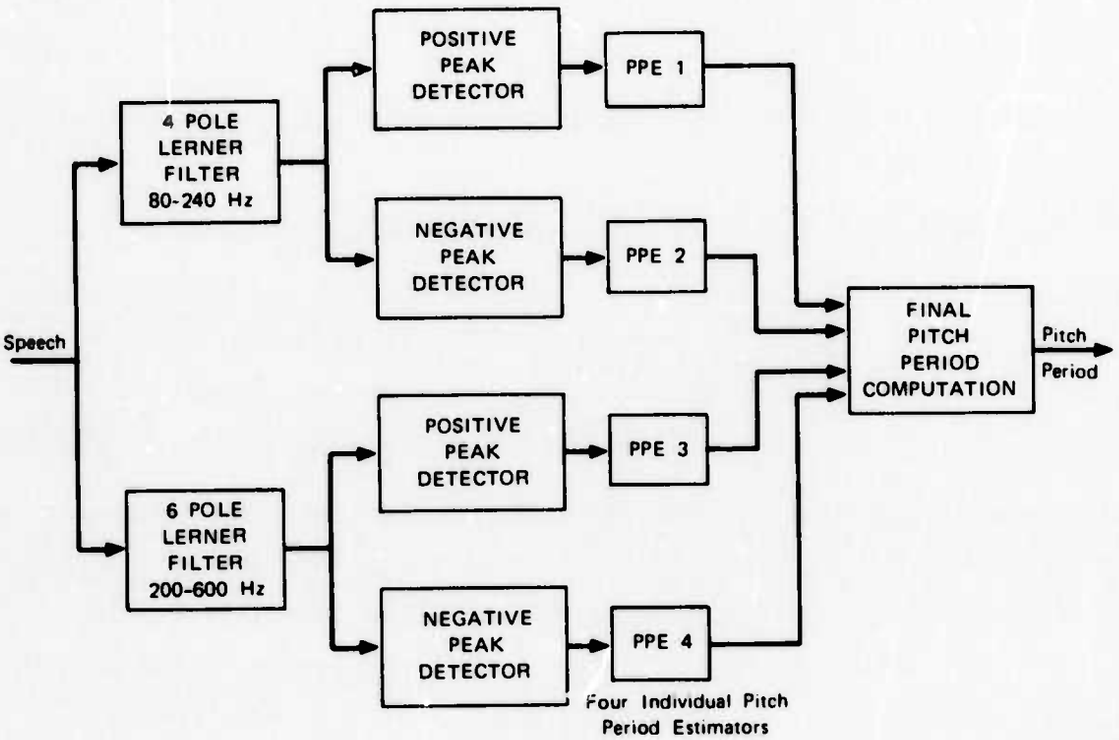
2. Time-Domain Pitch Extractor

In this section we describe a simple time-domain pitch extractor for use on the output of the two formant-isolation filters. By

"time domain" we mean that the pitch-pulse placement circuit operates directly on a time-domain waveform. A potential advantage of the time-domain approach is that it permits absolute as well as relative pitch-pulse placement. The former may be desirable if, for example, one wishes to do pitch-synchronous analysis. In addition, by using the time-domain approach, one can trade off response time for noise reduction more readily than by using the autocorrelation approach, for example. The Gold/Rabiner (G/R)[4] algorithm is a time-domain approach used in some vocoders. G/R algorithm provides relative rather than absolute pitch-pulse marks; however, it can be readily modified to provide absolute pitch-pulse marks.

Since the selected time-domain pitch extractor is similar to the G/R algorithm, we will describe the latter. (See Figure 13 for the salient features.) The speech signal is filtered into two bands, 80 to 240 Hz and 200 to 600 Hz, by Lerner filters. Positive and negative peak detectors work on the outputs of each filter, giving four pulse trains. Pulses occur at time points corresponding to potential pitch marks and have value equal to the absolute value of the speech waveform. Individual pitch-period estimators with detection circuits select candidate pitch pulses. The detection circuits have a variable blanking time, where no pulses are allowed, and a variable exponential decay. A pulse is selected whenever its value exceeds the value of the previous pitch pulse (for this circuit) times the exponential decay factor.

The individual pitch-pulse marks are used to calculate several pitch period measurements. As a result, the ability to place absolute pitch marks is lost at this point. These simple pitch-period measurements are then processed by a fairly complex processor to give the final pitch-period estimate. This processor is based on a majority vote concept so that occasional errors in one of the pitch-period measurements are corrected by the other estimates.



SOURCE: From Gold and Robiner[4].

SA-1526-38

FIGURE 13 BLOCK DIAGRAM OF THE PITCH-PERIOD ESTIMATION ALGORITHM

A modified version of the G/R algorithm using the formant-isolation concepts was developed at SRI (see Figure 14). Rather than using two formant-tracking filters, as described above, only one is employed. Due to difficulties in reliably tracking the second formant, we concluded that it was better to operate only on the first formant. To obtain the two filtered signals required by the G/R algorithm, we used a low-pass filter (Butterworth design) with a 3-dB bandwidth of 500 Hz. The F1 tracking filter (Lerner design) has a 3-dB bandwidth of 250 Hz. These filters were chosen to give a clean signal for pitch extraction from both male and female speakers. The selected filtering system represents a combination of the G/R and the formant-isolation concepts.

Positive and negative peak detectors are used on the filter outputs. The major peak estimators are similar to the G/R pitch-period estimators incorporating a blanking and run-down circuit, but have slightly different controls.

Experiments with the conventional G/R algorithm indicated many mistakes where half periods were indicated as periods due to the excessively quick run down. (Most of these were corrected by the logic in the final pitch-period computation.) The quick run down is required when the speech amplitude is decreasing (e.g., a vowel to consonant transition) to ensure picking up the next reduced amplitude peak. The characteristic of the run down was modified to yield fewer false half-period peaks. As a result, normal behavior was maintained for aperiodic marks, and periodic marks were detected more reliably.

The major change made in the G/R algorithm lies in the final pitch-period computation function. Instead of using peak locations to estimate the period, several <u>period</u> measures are used (1) to identify which filter has the best major peak and (2) to validate the peak as a

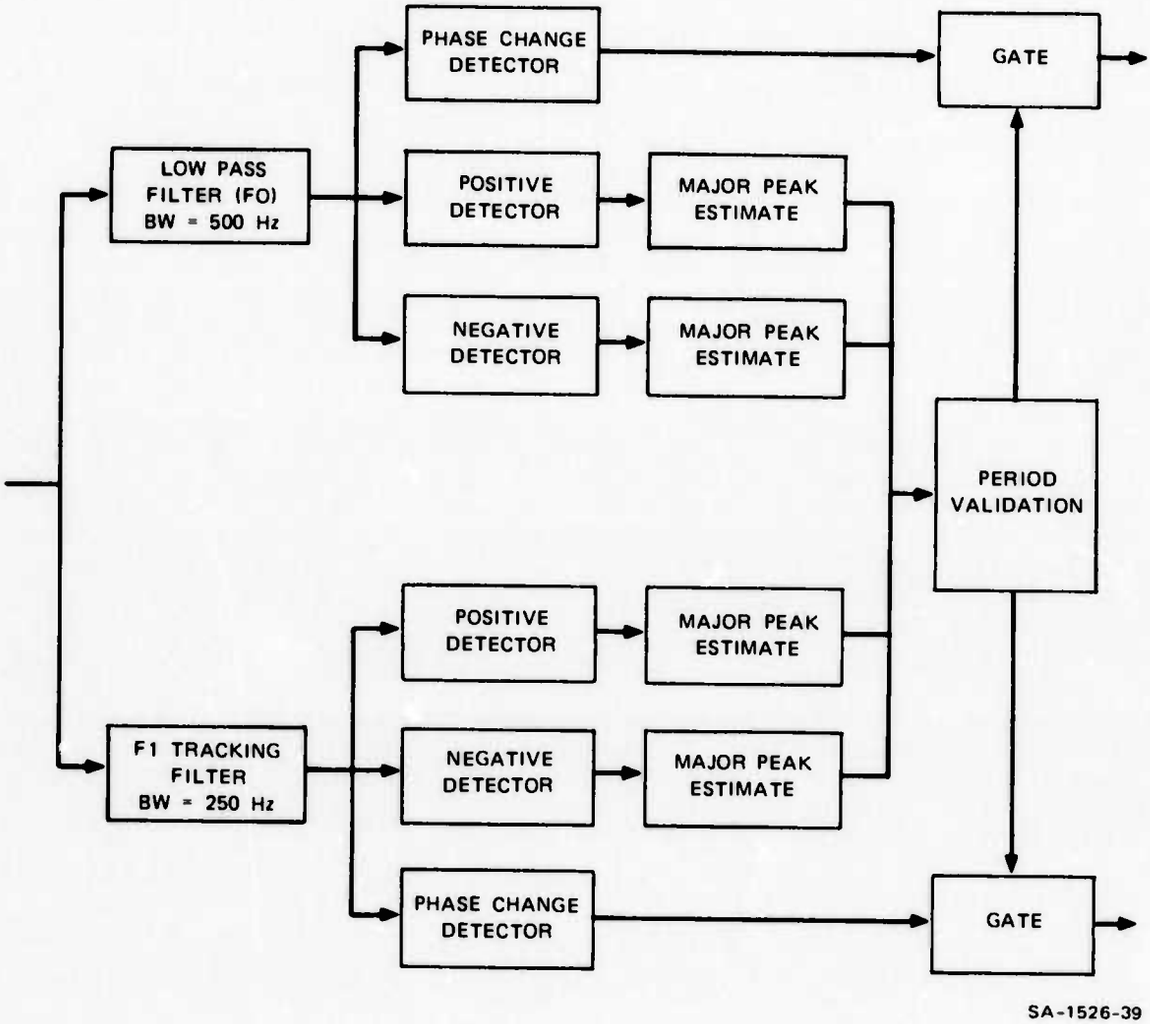


FIGURE 14 BLOCK DIAGRAM OF TIME DOMAIN PITCH EXTRACTOR

major peak. The pitch-mark estimate is then made in absolute time on the basis of the above information.

The types of time measurement used to detect start of excitation are illustrated in Figure 15. The top trace is the rms envelope of the sentence, "Add the sum to the product of the first three," for a female speaker. The middle trace is the output of the F1 tracking filter,

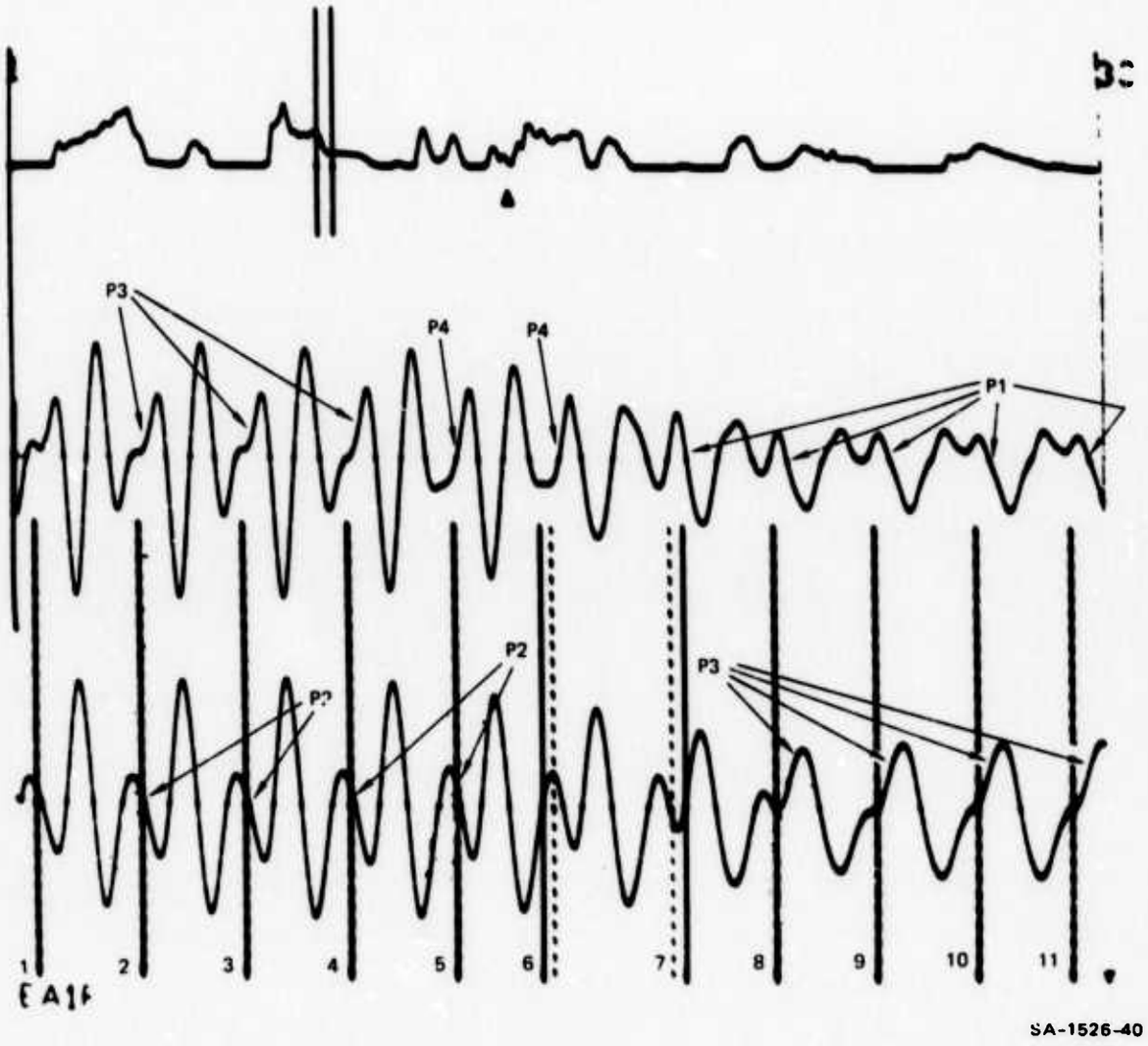


FIGURE 15 PITCH MARKS TYPES ILLUSTRATED ON SPEECH WAVEFORMS

and the bottom trace is the output of the low-pass filter; both traces are for the same short speech segment identified by the closely spaced lines in the top trace. This segment is the transition from the /a/ to the /m/ in "sum."

The types of peak marks employed are:

- P1--The zero crossing after a series of low amplitude peaks and before a major peak.
- P2--The zero before series of increasing peaks after a series of decreasing peaks.
- P3--A waveform discontinuity immediately preceding a major peak.
- P4--A region of low slope. (The exact mark location is determined so that the pitch periods change smoothly.)

These rules for pitch-mark location are given in their order of precedence. That is, a P1 mark is selected over a P2 mark, and so on. These rules were derived from a large number of human-aided pitch-marking experiments.

By inspecting all the major peak estimates and their absolute time marks, one is able to select the estimate, i.e., the channel, that gives the most consistent period estimate. Then additional filters (e.g., an F2 tracking filter, a feature detector, or major peak estimators) can be added if necessary.

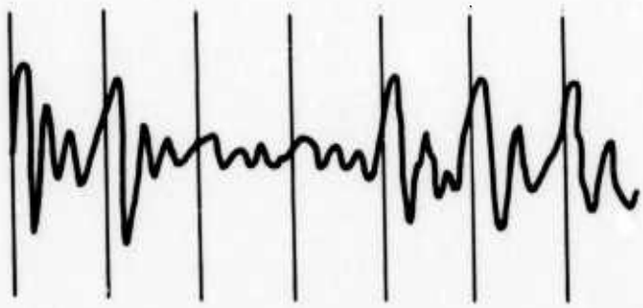
The algorithm outlined above (without the additional features) is presently running with a simulated LPC vocoder on SRI's PDP-10. There are significant periods of time when the correct pitch mark is made. However, it has been necessary to use human intervention to correct some errors. The quality of the synthetic speech needs further work if acceptable standards are to be met. Thus, a more complex algorithm appears to be required.

Filtering of the speech signal does facilitate pitch extraction; however, not all problems are solved by this process. Consider the word "unscrew" spoken by a male (RB); Figure 16(a) shows the input speech. Two low amplitude periods signal the transition from [n] to [r]. These periods appear to be voiced, and hand-placed marks are assigned to give a smooth period transition. Figure 16(b) shows the output of a Lerner bandpass filter tracking the first formant. The dashed marks in Figure 16(b) are the output of a G/R pitch-mark estimator based on positive peaks. The solid marks are the output of a similar pitch-mark estimator based on the negative peaks. Note that the negative marks precede the positive marks before the transition, but then a phase change occurs and the positive marks precede the negative marks. The G/R algorithm, which combines the estimator outputs, compensates for this phase change by simply estimating periods, i.e., the differences between marks from the same stimator, and combining results. However, for simple time domain (i.e., absolute-time placement of pitch marks) this is not possible. As a result, a discontinuity in pitch-pulse location and pitch period occurs. This is extremely noticeable. As shown in Figure 16(a), it is possible to hand-mark pitch pulses to avoid this problem. However, development of an automatic algorithm appears to be difficult.

3. Summary

We ceased research in the formant-isolation approach to concentrate our full efforts on the short-term memory, or residual-encoding, approach. At that time, it was clear that the formant-isolation approach had too many serious problems.

First, the basic concept of formant isolation attacks only one of the several problems of pitch extraction. Therefore, it should not be expected to work under all circumstances unless special precautions



(a) INPUT SPEECH WITH BEST HAND-MARKED PITCH PULSES

(b) OUTPUT OF FIRST-FORMANT TRACKING FILTER WITH POSITIVE (DASHED) AND NEGATIVE (SOLID) SETS OF PITCH MARKS

SA-1526-41

FIGURE 16 WAVEFORMS WITH SETS OF PITCH MARKS

are employed in the design of all its modules, e.g., the time-domain pitch extractors. This is a fundamental research problem in itself.

Second, the formant-isolation approach is complex (see Figure 11). It requires two additional filters (one of which is tunable),* a spectral envelope calculation (from the LPC parameters), and formant determination, in addition to the time-domain pitch markers and decision circuit.

Third, formant extraction is difficult. Unreliable extraction results for the higher formants; if these are needed to solve an ambiguity that the baseband and F1 cannot handle, difficulties may result. At present, the effect of serious formant errors is unknown.

Fourth, narrow-band formant-tracking filters provide poor time resolution. Consequently, in noisy environments, poor accuracy occurs, resulting in "rough" synthetic speech due to inaccurate location of the excitation pitch pulses.

The possible advantages of the formant-isolation approach were described earlier in some detail. The major intuitive advantage is that waveform simplification will result, thereby simplifying time-domain pitch extraction. However, it has yet to be demonstrated that the formant-isolation approach functions well where other pitch extractors, e.g., SIFT, fail. Formant isolation may indeed offer advantages, but no positive demonstration of this fact has been made.

The formant-isolation approach has proved extremely useful in assiting the process of hand marking pitch pulses, a process difficult even for an experienced pitch marker. Frequently several iterations

---

*Improved performance was obtained by extracting pitch from the baseband and from the first formant, rather than from the first and second formants.

are required before acceptable high quality synthesis results. Use of the formant-isolation system reduces the number of iterations and speeds the process of pitch marking. Thus, formant-isolation concepts have proved useful for experimentation and for development work.

Hand-marked pitch pulses have been developed for a data base on the SRI-AI PDP-10 computer system. Use of these pitch-mark sets with a 14-coefficient LPC synthesis at a 10-kHz sampling rate generates synthetic speech that is virtually indistinguishable from the original speech. Thus, a reference point (or rather a set of points) has been developed for automatic pitch-extraction algorithms. Their performance can be compared with that obtainable from the best hand-marked pitch-pulse locations. The comparison can be made on the basis of the (subjective) quality of the synthetic speech and on the basis of the (quantitative) rms time error between the locations of the hand-marked and the automatic algorithm-marked pitch pulses. This data base is available at SRI for use by anyone who desires to test the performance of his pitch extractor. The data base can be increased by applying hand-marking techniques (based on the formant-isolation concept) to the new speech data.

In summary, the real value of the formant-isolation concept is as a method of generating (with human assistance) in non-real-time an experimental set of very high quality pitch marks. These marks can then be used as a reference set for comparison with the results of more practical real-time pitch extractors.

III SHORT-TERM MEMORY APPROACH
(Residual-Excited Linear Prediction Vocoder)

A. Introduction

As an alternative to speech coders based on pitch-excited LPC, SRI is developing a residual-excited linear predictive vocoder (hereafter the SRI system is termed the RELP vocoder). The major difference between the two coding methods lies in the selection of the parameters characterizing the excitation signal that are to be transmitted over the channel.

For pitch-excited LPC the vocal tract, glottal flow, and radiation are represented by the prediction filter coefficients. Those coefficients are transmitted along with the information regarding excitation of speech, i.e., the fundamental frequency or pitch, F_o, the V/UV decision, and a gain, A, extracted from either the residual signal or the speech input. Pitch extraction is one of the most critical parts in LPC analysis, since the quality of synthesized speech is greatly affected by the reliability of the V/UV decision and by the accuracy of the location of the pitch pulses that drive the LPC synthesizer. Nevertheless, a pitch-extraction algorithm that is fully reliable and simple enough for hardware implementation is yet to be found, although research on pitch extraction has been pursued for over 30 years.

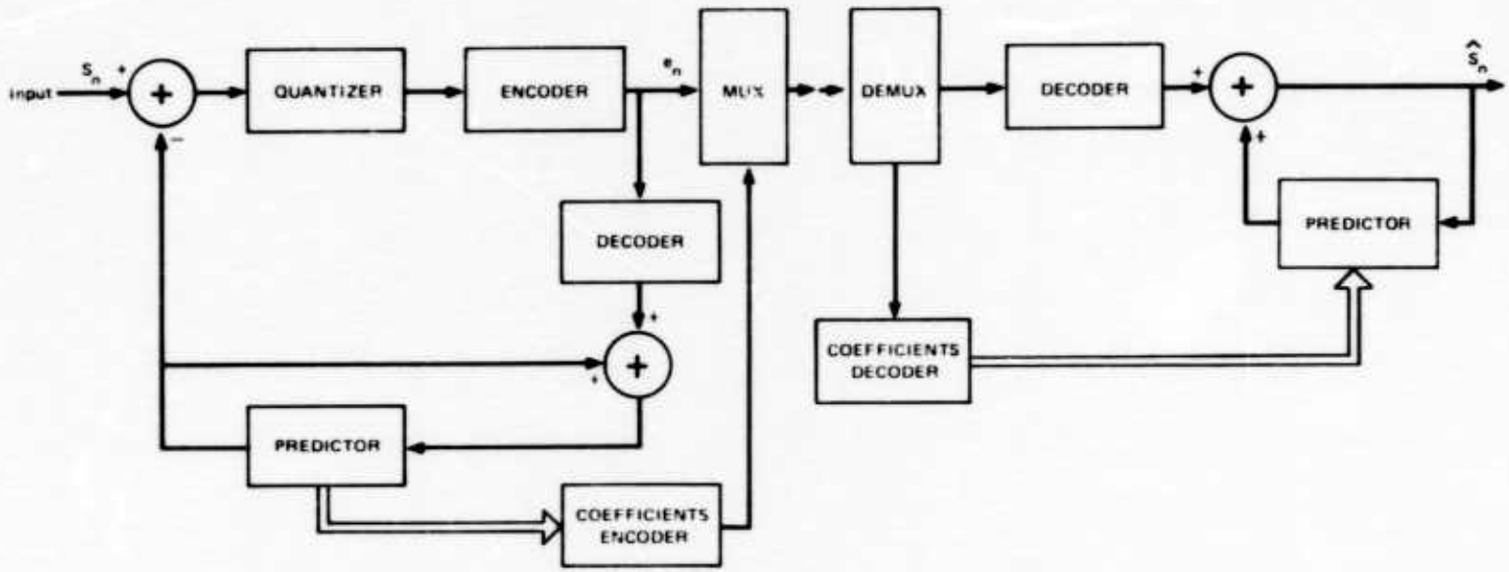
In a residual-excited linear predictive coder, the vocal tract is characterized in the same way as in a coder based on pitch-excited LPC. However, instead of the feature properties (F_o, V/UV, and A) of excitation being extracted and transmitted, the residual signal is encoded and transmitted, thus avoiding the difficult problem of pitch extraction. At the synthesizing end, the received residual is used instead of pitch pulses

as the excitation signal for driving the synthesizer. The SRI RELP system differs from other such systems in that the residual signal is low-pass filtered and then nonlinearly processed before being fed to the synthesizer. The low-pass filtering and nonlinear processing of the residual signal in the RELP system are, as we discuss later, similar in concept to the voice-excited vocoder (VEV).[16] The transmission rate is low (6 K to 9.6 K bits/s) compared with that of similar vocoders, yet very good quality results have been achieved with RELP under a variety of circumstances--some of which are most difficult. For example, RELP simulations have demonstrated good results with two simultaneous speakers. Consequently, RELP is viewed as a leading candidate for a practical low-rate (6 K to 9.6K bits/s) vocoder.

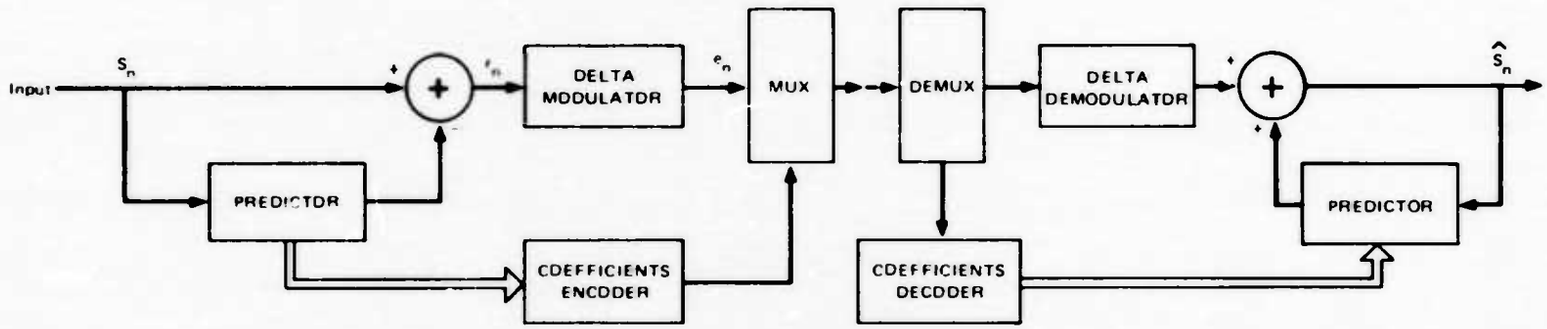
Before considering the RELP vocoder system, we briefly review existing LPC residual encoding methods. General discussion of the RELP system follows. We then discuss in detail each of the functional blocks of the RELP system and also the results of computer simulation. Finally, we consider the advantages of the RELP system and make a conclusion. This annual report accompanies an audio tape of various test utterances generated by the RELP vocoder simulated on an Interdata 70 minicomputer.

B. Review of LPC Residual Encoders

Basically, two different residual-encoding systems exist, their difference depending on the location of the encoder/quantizer and on how speech is predicted. One system puts the quantizer inside a linear-predictive loop, and the prediction of speech is based on previously reconstructed speech samples and the error signal. The other system puts the quantizer outside the linear-predictive loop, and the prediction is based on the previous input speech samples. The two systems are shown in Figure 17. The first encoder uses DPCM [Figure 17(a)], which has been



(a) DPCM



(b) DM

SA-1526-4

FIGURE 17 RESIDUAL ENCODING SYSTEMS

studied by McDonald,[17] Melsa et al.,[18] and Atal and Schroeder;[2] the second encoder uses ADM,* which has been investigated by Dunn.[19]

McDonald's approach uses straight DPCM with both a predictor and a quantizer fixed. The approach of Melsa et al. uses adaptive DPCM, where the residual signal is encoded with a variable multilevel quantizer and transmitted to the synthesizer. A significant feature of the latter approach is that prediction coefficients are not transmitted but rather are generated from the transmitted residual and the synthesized speech. The typical number of prediction coefficients is eight; the sampling rate is 6.6 kHz; the transmission data rate in adaptive DPCM is 16 kHz. Although the quality of synthesized speech is reasonably good at the above values, the disadvantage of the method is that the data rate is relatively high compared with that of a pitch-excited LPC.

Atal and Schroeder have used a more elaborate predictor by taking account of the characteristics of speech sounds but have used a quantizer with one bit to reduce the bit rate. In their system a delay parameter corresponding to the pitch period calculated by the minimum-mean-square-error process, eight suboptimized predictor coefficients, and a gain are transmitted, together with the residual signal encoded by the one-bit quantizer with adjustable step size. They claim that their encoding method accomplishes reduction of signal redundancies with a low bit rate (about 10K bits/s) and that the quality of the synthesized speech is comparable to that of log PCM speech encoded at 6 bits/sample. However, it is not clear whether the prediction filter suboptimized on the basis of the optimum delay parameter can be comparable, in terms of the mean-square

---

*Actually, in the second encoding scheme the residual signal generated from LPC analysis can be encoded also by PCM or DPCM. But, since ADM among the three yields the best SNR in the low range of transmission data rate, our discussion is limited to ADM.

prediction error, to the LPC filter obtained by straight optimization. Inaccuracy of the delay parameter or pitch can destroy the optimum of the filter coefficients. Furthermore, it is doubtful whether the accuracy of the delay parameter, which is very critical in synthesis, can be guaranteed with the optimization process of Atal and Schroeder in some adverse circumstances, such as phase change from one pitch period to another owing to the presence of high-frequency signal components. Because of the requirement of a high data rate in the adaptive DPCM of Melsa et al. and the necessity of having the delay parameter for the pitch period in the system of Atal and Schroeder, we do not consider those two systems here.

Dunn has considered encoding the residual signal in a different way from the above system [see Figure 17(b)]. LPC residual signal is generated by a feed-forward LPC analyzer and is encoded by delta modulation. It has been reported that the resulting quality of synthesized speech is not as good as that of ordinary PCM coding using the data rate of 50 K bits/s but is comparable to that of ADM coding with the rate of 20K bits/s. The cause of the inferior quality of Dunn's synthesized speech is believed to be the low sampling rate (8 kHz) of delta modulation in encoding a residual signal with a bandwidth 3.4 kHz.

After some initial consideration, SRI selected the feedforward method for the following reasons. First, it is somewhat simpler; no consideration of stability is required for the analyzer. Second, the feedforward approach is most compatible with the VEV concept; i.e., a spectral flattener operating on a transmitted portion of the voice baseband can be used. By contrast, the operation of a feedback analyzer would be greatly complicated by the presence of the nonlinearities of the spectral flattener occurring in the feedback loop. In other words, the feedback approach is basically designed for waveform matching while the feedforward

approach permits use of spectral matching techniques, such as those employed in the VEV, that permit a greater bandwidth compression of the residual signal. Third, and most importantly, the feedforward approach permits great flexibility, since it is inherently a modular concept. If good pitch extractors become available in the future, then the residual encoder can be replaced and a lower rate system will result. Another possibility is that pitch extractors would be employed in favorable acoustic environments, while the residual encoder would be employed for noisy environments or for multiple speakers.

C. RELP Vocoder System

1. General

The motive for conceiving the SRI RELP vocoder system shown in Figure 18 was to avoid the difficult process of pitch extraction and instead to transmit the residual as an excitation signal with low bit rate in linear predictive coding of speech. The region of our interest in the total transmission bit rate is that between 6K and 9.6K bits/s, which is below the data rate of the other residual-excited predictive coders discussed in the preceding section but about 3K to 6K bits/s above that of a pitch-excited LPC vocoder. The necessity of having an additional 3K to 6K bits/s in the RELP vocoder system may be interpreted as the trade-off for not having the pitch extraction required in a pitch-excited coder. If one transmits a signal by delta modulation, usually the sampling rate must be several times the highest-frequency component of the signal. Hence, for a residual signal generated from LPC analysis of a speech band-limited to 4 kHz, one would need a sampling rate or delta modulation bit rate of at least 20 kHz in coding the signal by delta modulation. This sampling rate or delta modulation bit rate is well above the goal of achieving the bit rate between 6K and 9.6K bits/s.

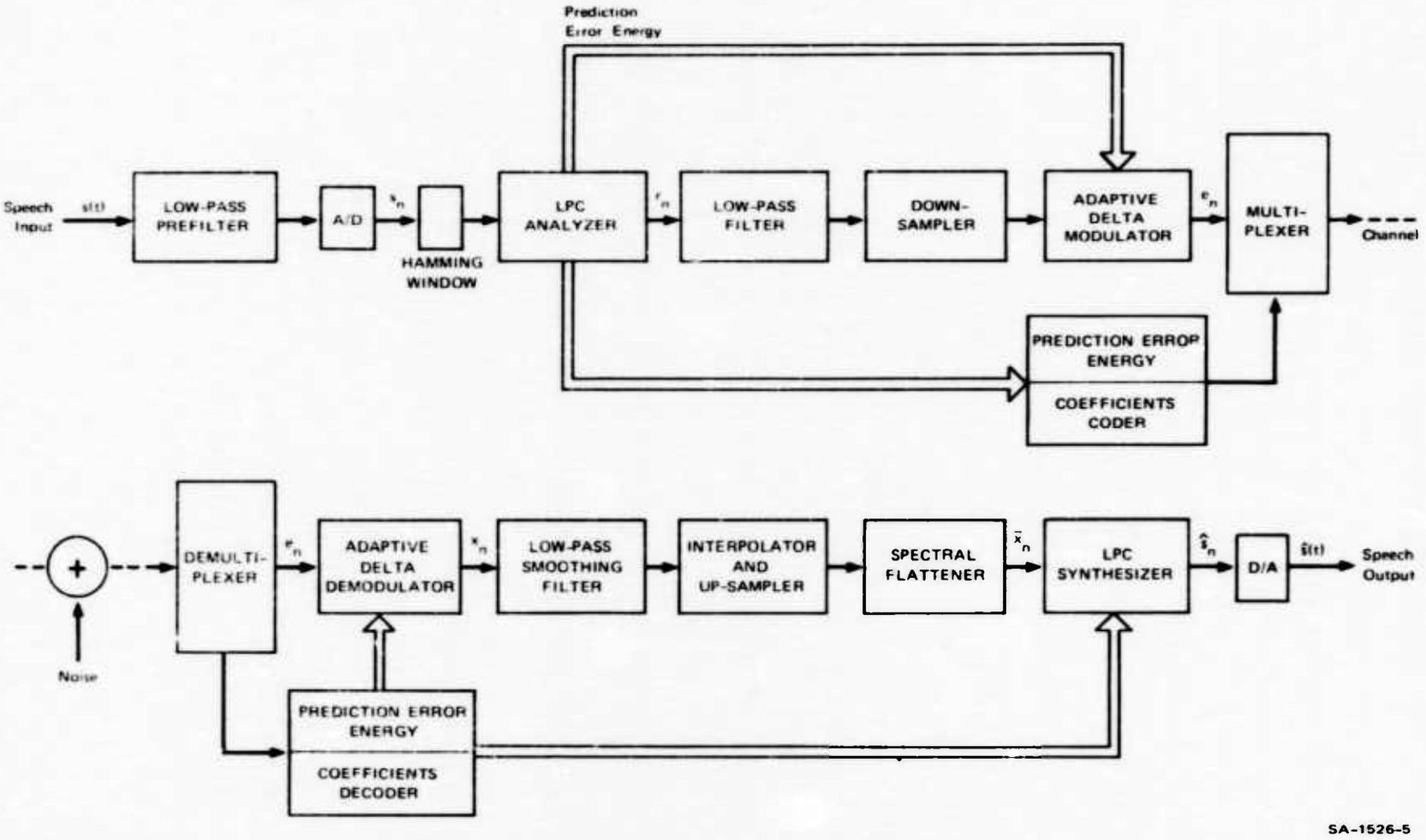


FIGURE 18 RESIDUAL-EXCITED LINEAR PREDICTION VOCODER

In the RELP vocoder system we reduce the bandwidth of the residual by low-pass filtering it with a cutoff frequency of 800 Hz.* This enables us to have an ADM sampling rate below 7 kHz without excessive quantizing distortions. Low-pass filtering of the ADM input signal reduces the whitened effect of the LPC residual and consequently results in a waveform that resembles a speech wave more than a very spiky impulsive wave. Since the whitened LPC residual with an impulsive waveform is very difficult to code by delta modulation, this "taming" effect is important in reducing quantizing noise in ADM coding.

After the residual signal has been low-pass filtered, it is down sampled before coding, since the original residual is generated from LPC analysis of speech signal sampled at the rate of 10 kHz. If a higher rate is desired for some reason, one may skip the down-sampler. The low-passed residual may be coded by any digital encoding method, e.g., PCM, differential PCM, or delta modulation. Since delta modulation has been shown to yield the least quantizing noise in encoding a signal at a low data rate, we chose ADM as our encoding method. The ADM-coded residual is multiplexed with coded prediction coefficients and prediction energy and then transmitted over a channel to the receiver.

If the low-passed residual has been down sampled at the transmitter, the decoded residual at the receiver must be up sampled by interpolation before being fed into the LPC synthesizer, so that the sampling rate at the synthesizer will be the same as that at the analyzer. The signal is then passed through a spectral flattener to recover the high-frequency components of the original residual. The spectral flattening is done by a nonlinear distortion processing using an asymmetrical linear full-wave rectifier. The spectrally flattened signal is then mixed with

---

*We have also used a 400-Hz cutoff frequency.

random noise regardless of whether the input speech is voiced or unvoiced. The amount of random white noise mixed with the excitation signal is, of course, varied in different phonemes by using the prediction error energy computed at the LPC analyzer. The purpose of adding random noise is to increase the high frequency energy of the excitation signal and consequently to improve the synthetic speech quality. After the received signal has been thus conditioned, it is used as an excitation signal to drive the LPC synthesizer. The digital signal output of the synthesizer is finally converted into an analog signal by a digital-to-analog (D/A) converter to obtain the synthesizer speech.

One might note that the general concept of the RELP vocoder system is similar to that of the VEV originally developed by Schroeder and David.[16] In the VEV the vocal tract is characterized by a set of bandpass filters encompassing the frequency range of speech signal, and an unprocessed baseband of the original speech with its upper frequency limited to 900 Hz is transmitted as an excitation signal to the synthesizer. At the synthesizing end this baseband is passed through a nonlinear distortion process as in the RELP system, to spectrally flatten and broaden it. The resulting signal is then used as the source of excitation to drive the vocoder channel filter bank. Because the excitation signal has been derived from the real speech band, it inherently preserved the pitch information, i.e., the fundamental frequency and V/UV decision, which are critical for good-quality synthesized speech. Because of the preservation of the pitch information in the excitation signal, the quality of the synthesized speech has been found superior to that of a channel vocoder excited by pitch pulses. The naturalness of the VEV speech sound is preserved, while the synthesized speech of pitch-excited channel and formant vocoders has a mechanical quality. It should be noted that, although VEV has a considerably higher data rate than a

pitch-excited channel vocoder because of the transmission of the baseband signal, it still yields a bandwidth compression of about three to one.

In comparing our RELP vocoder system with the VEV, one may note that the linear predictive filter for characterization of the vocal tract corresponds to the channel filter bank in the VEV, and the low-passed LPC residual signal for excitation of the synthesizer in the RELP vocoder corresponds to the baseband speech signal in the VEV. It is well known that the very flexible transfer function of a linear predictive filter permits better matching of the envelope of the speech short-term power spectrum than does a bank of fixed bandpass filters. Consequently, one may expect the RELP vocoder to yield synthesized speech superior in quality to that of the VEV, under similar operating conditions.

2. Detailed Discussion and Computer Simulation

The RELP system is now discussed in detail in order of the signal flow. The results of computer simulation are considered in appropriate parts of the ensuing subsections. (The reader can also find a summary of the parameters used in the RELP simulation in Subsection 2-g and the computer flow charts in Figures 38 through 40.)

a. Preprocessing of Speech Signal

Since the sampling rate of a signal and the number of LPC coefficients are related to the bandwidth of the input signal, the speech input must be band-limited before digitization to obtain the desired results. In our simulation the speech input has been low-pass filtered by an analog Butterworth filter with a cutoff frequency of 3.2 or 4 kHz and skirt decay of 85 dB per octave. The cutoff characteristic of the filter was made sharp to minimize the aliasing problem. The low-passed signal has been sampled at the rate of 6.8 kHz for the signal with a cutoff

frequency of 3.2 kHz and at 10 kHz for the signal with a cutoff frequency of 4 kHz by a 12-bit analog-to-digital (A/D) converter and stored in data blocks for LPC analysis.

b. LPC Analysis

Two distinctly different LPC analysis methods have emerged since research on linear predictive coding of speech began: the covariance method due to Atal and Hanauer[9] and the autocorrelation method due to Itakura and Saito[20] and Markel.[21] In our RELP system the second method has been used, primarily because it is computationally efficient and less prone to instability of the synthesizing filter. Since these two methods have been discussed in detail in the literature,[9,21,22] we summarize here only the autocorrelation method, which is the method pertinent to our RELP computer simulation.

In the autocorrelation method of LPC analysis, the pre-processed speech samples are windowed by the Hamming window,

$$W_n = (0.54 - 0.45 \cos 2\pi n)/N \quad , \tag{9}$$

to generate

$$s_n = \text{windowed speech samples}, \ 0 \leq n \leq N-1$$

$$= 0, \text{ otherwise} \quad , \tag{10}$$

where N is the window length. For our simulation the length, N, of the Hamming window was made 256 sample points, and the length of an analysis block was made 200 sample points. Hence the LPC analysis of a block, say B_k, has been done with 256 windowed speech samples made of 28 samples of the previous block B_{k-1}, 200 samples of B_k, and 28 samples of the next block. For analysis of the next block, B_{k+1}, the window is moved

by 200 samples. This overlapped windowing gives a smoothing effect in LPC analysis, i.e., avoids abrupt change in LPC coefficients between analysis blocks; furthermore, overlapping (or weighting them very slightly) avoids missing data samples that fall in the window nulls. Consequently, the quality of the synthesized speech is better than that obtained by nonoverlapped windowing.

When the windowed speech samples have been generated, a sampled speech signal, s(nT), at discrete time, t = nT, is predicted by the past p samples as

$$\tilde{s}_n = \sum_{k=1}^{p} a_k s_{n-k} \quad , \tag{11}$$

where $\tilde{s}$ is the predicted value of s(nT) or s_n, and $\{a_k\}$ is a set of real constants that represent the predictor coefficients. The predictor coefficients are determined by a minimum-mean-square-error process. The error between the predicted and real speech samples is given by

$$e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k} \tag{12}$$

or in Z domain

$$E(z) = A(z)\, S(z) \quad , \tag{13}$$

where

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \quad , \tag{14}$$

and E(z) and S(z) are z-transforms of e_n and s_n, respectively. The rms energy is then minimized by the discrete Wiener process[23] over all n or time. This results in the autocorrelation equation

$$\sum_{k=1}^{P} a_k R_{|i-k|} = R_i, \qquad i = 1,2,\ldots p \quad , \tag{15}$$

and the minimized energy

$$E_{min} = R_o - \sum_{k=1}^{p} a_k R_k \quad , \tag{16}$$

where

$$R_i = \sum_{n=o}^{N-1-|i|} s_n s_n + |i|, \qquad R_i = R_{-i} \tag{17}$$

and

$$R_o = \sum_{n=o}^{N-1} s_n^2 \quad . \tag{18}$$

The autocorrelation coefficient $R_{|i-k|}$ in Eq. (15) may be expressed in a matrix form:

$$R_{|i-k|} = \begin{bmatrix} r_0 & r_1 & r_2 & \cdots & r_{p-1} \\ r_1 & r_0 & r_1 & \cdots & r_{p-2} \\ \vdots & \vdots & \ddots & & \\ r_{p-1} & r_{p-2} & r_{p-3} & \cdots & r_0 \end{bmatrix} \quad , \tag{19}$$

which is a Toeplitz matrix. To solve Eq. (15) for the prediction coefficients, the matrix $R_{|i-k|}$ is inverted by Robinson's modified method of Levinson's algorithm.[23,24] The stability of the recursive synthesizing filter,

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \quad , \tag{20}$$

is guaranteed, at least theoretically, when the matrix $R_{|i-k|}$ is Toeplitz.[12]

So far our discussion on the autocorrelation method of LPC analysis has been rather general. It has been found that in the RELP vocoder it is advantageous to preemphasize or difference speech samples before LPC analysis, i.e.,

$$s'_n = s_n - cs_{n-1} \tag{21}$$

where s'_n is a differenced sample and c is a constant. The block diagram of the LPC analyzer with a differencer is shown in Figure 19. Note that the differencer is applied only to the input to the LPC analyzer. The constant, c, of the differencer is chosen such that the breaking point

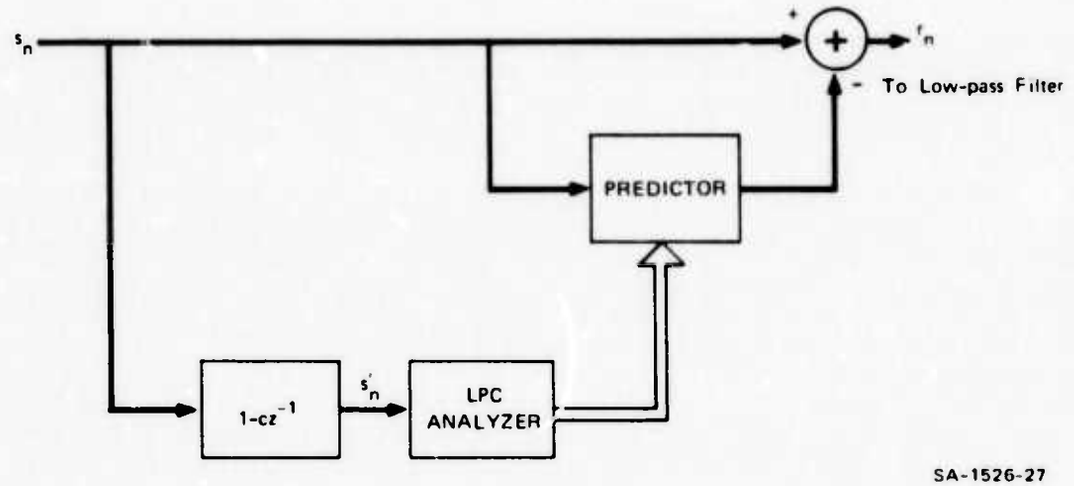


FIGURE 19 LPC ANALYZER WITH A DIFFERENCER

of the differencer occurs at the same value as for the low-pass filter of the residual signal.

The major advantage of preemphasizing speech samples before LPC analysis in the RELP vocoder is that it increases the high frequency content of the prediction filter and thus offsets the effect of low-pass filtering of the excitation signal. Another important advantage is that preemphasis reduces the spectral dynamic range and, consequently, coding of the LPC coefficients with preemphasis results in more accurate quantization.[25] The effect of preemphasis on the synthetic speech quality will be discussed in Subsection 2-g. Detailed analysis of the effect of preemphasis in the RELP system is being made.

The number of filter coefficients can be varied depending on the specific application and the input signal bandwidth. In an application of the prediction filter to a speech signal band-limited to 4 kHz, the typical number of coefficients is about 12. The spectrum of the speech band-limited to 4 kHz has at least three formants. Since the poles of the prediction filter represent the formants of the vocal tract, and they occur in complex conjugate pairs, the number of coefficients should in this case be at least six for adequate spectral matching.

However, if the one-step prediction error or residual signal generated by a prediction filter is transmitted to the LPC synthesizer and used as an excitation signal of the synthesizer, the number, p, of filter coefficients could theoretically be any value, since the inverse operation of Eq. (13) always holds:

$$S(z) = \frac{E(z)}{A(z)} \quad . \tag{22}$$

The larger p is, the less the residual signal should include the formant structure of input speech, or vice versa. In the

extreme case of $p \rightarrow \infty$, E(z) will be purely white, indicating that E(z) has only the excitation information. This inverse property is one of the reasons why most residual-excited LPC coders have fewer coefficients transmitted than a pitch-excited LPC vocoder does. The residual signals generated by a prediction filter with different numbers of coefficients are shown in Figure 20. Although the waveforms with ten and six coefficients do not seem to differ, the residual generated with fourteen coefficients is much whiter than the residual with six or ten coefficients.

Note that although a residual signal is transmitted and used as the excitation signal of the synthesizer in our RELP vocoder, the signal is not the above-mentioned residual, but rather is a low-pass-filtered one having mostly the pitch information. Therefore, unlike other residual-excited coders, the number of filter coefficients of the RELP vocoder must be comparable to that of a pitch-excited LPC vocoder to obtain speech of good quality.

The quality of synthesized speech was tested with different numbers of coefficients in RELP simulation. The quality was good with 14 coefficients, and little degradation of quality resulted from lowering the number to ten for a sampling rate of 6.8 kHz. One could detect, however, the difference of quality between eight and ten. We have concluded from this experiment that the optimum number of coefficients is ten for a sampling rate of 6.8 kHz and twelve for a sampling rate of 10 kHz.

In addition to computing the predictive coefficients and generating the residual signal, the LPC analyzer computes the energy of the residual signal in each analysis block. This is transmitted to the receiver along with the coefficients and the residual signal. The residual energy is, as we discuss later, used to provide information regarding syllabic companding for the ADM encoder and decoder and gain control of the excitation signal of the LPC synthesizer.

(a) 14 COEFFICIENTS

(b) 10 COEFFICIENTS

(c) 6 COEFFICIENTS

SA-1526-44

FIGURE 20 LPC RESIDUAL SIGNALS OF /o/ IN "OAK" GENERATED FROM A PREDICTION FILTER WITH DIFFERENT NUMBERS OF FILTER COEFFICIENTS

The predictive coefficients may be transmitted directly to the synthesizer or can be converted first to a set of coefficients called reflection coefficients $\{k_i\}$ and then be transmitted after coding. Since coding of reflection coefficients requires fewer bits and simplifies the stability check of the synthesizing filter, transmission of these parameters is generally preferred.*

The reflection coefficient parameters $\{k_i\}$ are obtained in the process of solving for the linear predictive coefficients by Levinson's method. The reflection coefficients are given by $k_i = a_{ii}$ for i = 1 to p where $\{a_{ii}\}$ represent the major diagonal elements of the triangular matrix developed in the Levinson solution of the autocorrelation equations. The recursive equations of the Levinson algorithm are presented in Figure 21. The expanding triangular matrix associated with these recursive equations is also shown. In this figure, the diagonal elements $\{a_{ii}\}$ correspond to the recursive equation quantities $\{a_i^{(i)}\}$. The major point is that the reflection coefficients can be obtained readily. In fact, the reflection coefficients are no more difficult to obtain than the conventional linear predictive coefficients.

c. Low-Pass Filter and Down Sampler

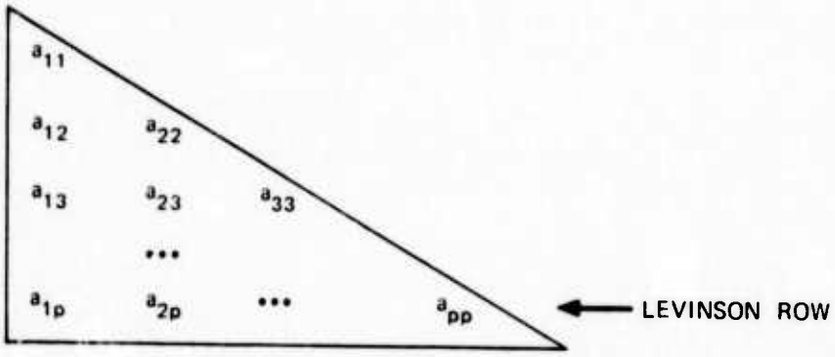
As mentioned in the preceding section, the purpose of the low-pass filtering of the residual signal is to compress the bandwidth of the signal and consequently to "tame" the whitened effect of the LPC residual before ADM coding. Figure 22 shows the original residual of /o/ in "oak" generated from LPC analysis with ten coefficients, the

---

*The values of the reflection coefficients are nonuniformly distributed over the interval [-1,1]. The necessary and sufficient condition for the synthesizing filter to be stable is that $|k_i| < 1$ for i = 1 to p. (See Markel and Gray.[25])

$$\underline{a}_{n-1}^{(n)} = \underline{a}_{n-1}^{(n-1)} - a_n^{(n)} \left[\underline{a}_{n-1}^{(n-1)} \right]^{\cdot}$$

$$a_n^{(n)} = \left[R_n - \sum_{k=1}^{n-1} R_{n-k}\, a_k^{(n-1)} \right] \Big/ \left[R_o - \sum_{k=1}^{n-1} R_k\, a_k^{(n-1)} \right]$$



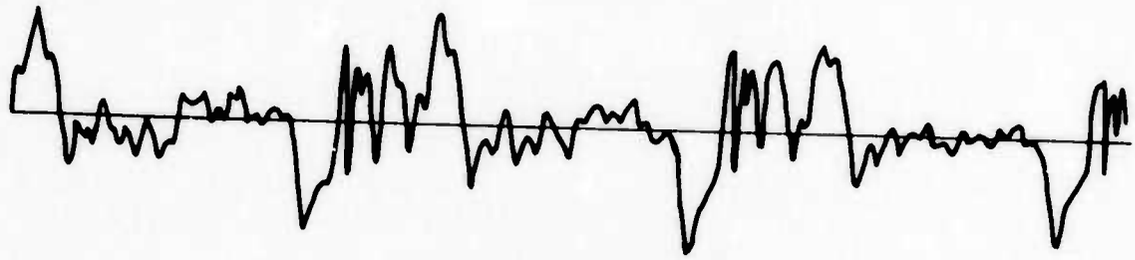
DEFINE $K_i = a_{ii}$

SA-1526-55

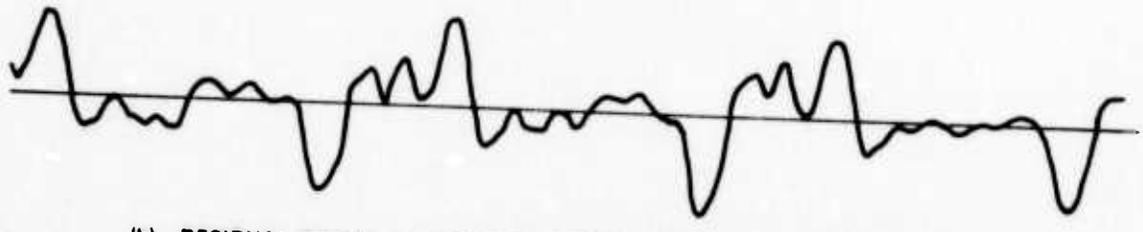
FIGURE 21 RECURSIVE EXPANSION TRIANGLE AND EQUATIONS

low-pass filtered residual with a cutoff frequency of 800 Hz, and the filtered residual with a cutoff frequency of 400 Hz. The low-pass filter used was a four-pole Butterworth filter whose skirt decayed 24 dB per octave. One can see the dramatic effect of the low-pass filtering of the residual in the figure. The effect of using different cutoff frequencies is discussed later when we consider ADM simulation.

The bandwidth of the low-pass filter must cover the whole range of the fundamental frequency of speech, 50 Hz to 450 Hz, so as to recover at the receiver the high-frequency harmonics, which have been filtered out at the transmitter. In the case of the telephone-line speech in which the lower 300 Hz is missing, one must have either the fundamental frequency or two adjacent harmonics to recover the high-frequency harmonics.[26] This means that in the latter case the low-pass filter should



(a) ORIGINAL RESIDUAL SIGNAL



(b) RESIDUAL SIGNAL LOW-PASS FILTERED WITH CUTOFF FREQUENCY 800 Hz

(c) RESIDUAL SIGNAL LOW-PASS FILTERED WITH CUTOFF FREQUENCY 400 Hz

SA-1526-45

FIGURE 22 LPC RESIDUAL SIGNALS OF /o/ IN "OAK"

be band-limited to 900 Hz. In our simulation the cutoff frequency chosen was 400 Hz, with 800 Hz for the telephone-line speech. Those two frequencies are actually a little low to cover the whole spectrum of the fundamental frequency, but they should be adequate for military communications.

When the residual signal has been low-pass filtered, the signal is down sampled to reduce the bit rate of the ADM. Note that the sampling rate of the ADM input corresponds to the transmission bit rate, since the ADM has a one-bit quantizer. In our simulation we have taken every other residual sample as the input to the ADM system. Hence, for a speech signal sampled at 10 kHz the ADM bit rate for transmission of the residual signal is 5K bits/s, and for a speech signal sampled at 6.8 kHz the transmission bit rate is 3.4K bits/s. The sampling rate of ADM usually must have at least several times the input bandwidth. Therefore, the ADM sampling rate of 5 kHz should be adequate for the input signal with the bandwidth of 2.00 Hz and is enough for the signal with the bandwidth of 400 Hz. If one wishes to increase the transmission rate for better quality of the synthetic speech, the down sampler may be skipped. In this case, the ADM transmission rate is the same as the sampling rate of the input speech.

d. Adaptive Delta Modulation

1) General

Since the invention of delta modulation (DM) by F. de Jager in 1952,[27] DM has gained in popularity as a simple, effective method of A/D conversion. The process of DM is simpler and possibly cheaper than the processes of PCM and DPCM. In spite of these advantages, DM did not have initial wide acceptance and was not considered competitive with PCM or DPCM. This lack of acceptance may have had two bases:

(1) DM was believed to require greater bandwidth, and (2) the dynamic range of linear DM,* which until recently was primarily used, is severely limited. This limitation of the dynamic range of DM results from the two types of inherent quantizing noises (see Figure 23): the granular noise produced by the finite step size of the encoder and the slope overload noise introduced when the slope of the input signal is greater than DM can follow.

In fact, the difficulty associated with a greater bandwidth (a higher bit rate) for DM does not generally have a valid basis. For a signal having most of its power at high frequencies, standard PCM outperforms DM in terms of signal-to-noise ratio. However, DM yields a larger SNR for a signal having most of its power at low frequencies. Speech signals tend to have most of their power at low frequencies, and to some extent this characteristic applies to the filtered residual. One of the reasons we chose DM for coding the low-passed residual is its ability to provide modest output SNRs with minimum transmitted bit rate.

The problem of the narrow dynamic range of linear DM can be overcome by making the quantizer step size adaptive, i.e., by varying the step size according to the magnitude or slope of the input signal. This is discussed along with the general principle of ADM with hybrid companding that we have used.

2) Theory of ADM with Hybrid Companding

Our ADM encoder and decoder are shown in Figure 24. Delta modulation may be regarded as the simplest form of differential PCM with a one-bit quantizer. The signal to be transmitted is periodically

---

*In linear DM, the quantizer step size is fixed.

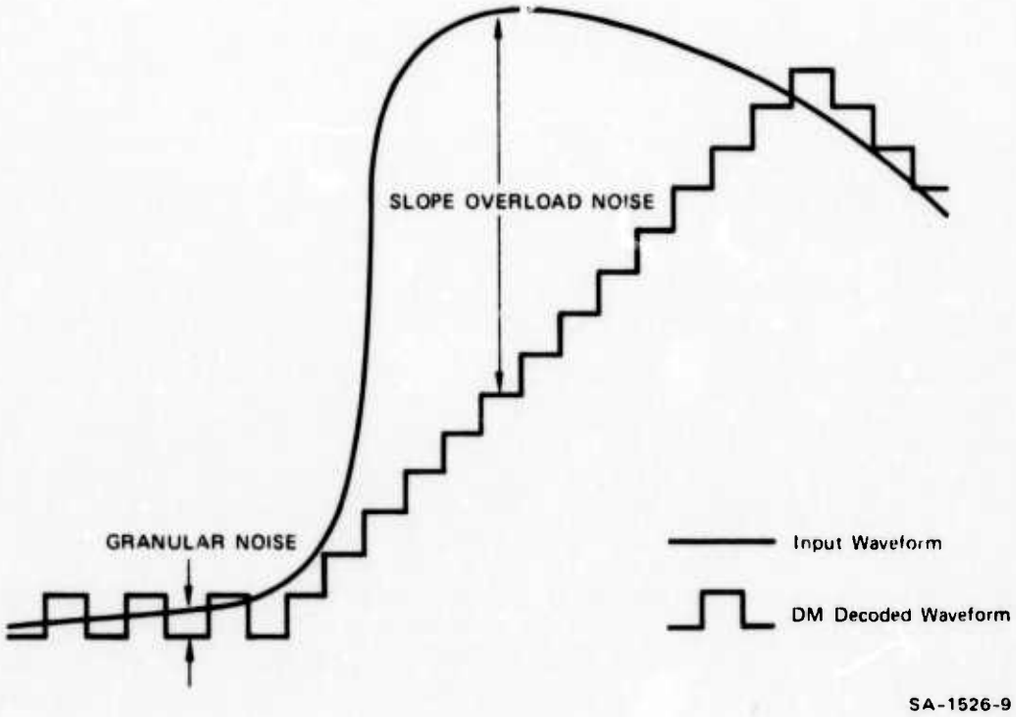


FIGURE 23 DM QUANTIZATION NOISES

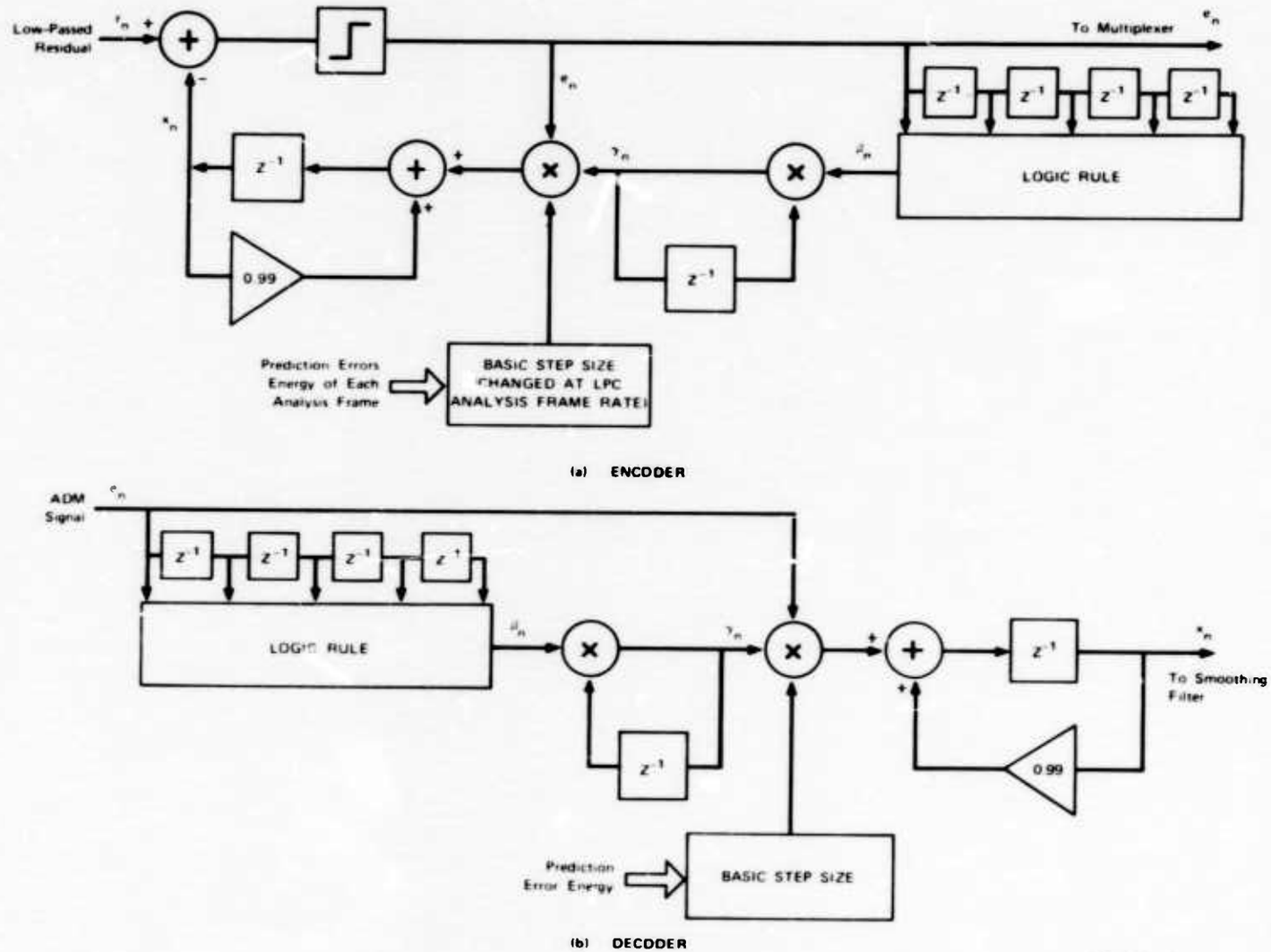


FIGURE 24 ADAPTIVE DELTA MODULATOR

sampled and compared with its estimated value, which is obtained by increasing or decreasing the previous estimate at each sampling time by one step size, depending on the sign of the difference between the signal and the estimate. The sign information, one bit per sample, is transmitted over a binary channel to the receiver, and these sign bits are used to construct the estimate of the original signal at the receiver.

Hence, given the sampled signal r_n in our ADM coder, sign bits are generated as

$$e_n = \text{sgn}\ (r_n - x_n) \quad , \tag{23}$$

with

$$x_n = x_{n-1} + e_n \Delta_n$$

$$\Delta_n = \gamma_n \Delta_{n-1}$$

$$\gamma_n = \beta_n \beta_{n-1}$$

$$\beta_n = f\ (e_n, e_{n-1}, e_{n-2}, e_{n-3}, e_{n-4}) \quad ,$$

where Δ_n is the n_{th}-step size, and β_n is a multiplication factor. Note that the basic step size, Δ_o, of the quantizer is obtained by

$$\Delta_o = \alpha E \quad , \tag{24}$$

where α is a scale factor, and E is one-step prediction error energy in an analysis block. Therefore, the ADM step size, Δ_n, is actually a function of the prediction energy, E; the multiplication factor, β_n; and the previous step size, Δ_{n-1}. The factor β_n depends on the present and previous four sign bits.

The ADM decoder at the receiver is exactly the same as the feedback paths of the ADM encoder. The estimate of the input signal is constructed on the basis of the received sign bits, the prediction error energy, and the logic rule. The feedback path has a gain with a value less than one to minimize the effect of transmission errors.

It is noteworthy that our ADM companding of the step size is hybrid, i.e., both syllabic and instantaneous companding schemes are used. The one-step prediction error energy provides the information of long-term step size at syllabic rate. The logic based on the five consecutive sign bits makes the ADM quantizer step size instantaneously compand. This hybrid companding is unlike other ADM step companding algorithms, in which companding is either syllabic or instantaneous.[28-33] Using hybrid companding in ADM should be advantageous, particularly for transmission of speech signal or its residual. There is a large difference in the dynamic range of speech signals, in general, and also between V/UV signals and among the different phonemes. Consequently, if one uses instantaneous companding and thus fixes the basic step size, the ADM system that works well for voiced signal may yield unacceptable quantization noise for unvoiced signal, or vice versa. The same is true for different phonemes. However, if one uses hybrid companding, he should not encounter the above difficulty, since the basic step size (which is, in fact, one-step prediction error energy) is transmitted at a syllabic rate or LPC analysis frame rate. Of course, hybrid companding necessitates a higher data rate or greater system complexity, or both, compared with instantaneous or syllabic companding. But the increase of data rate and complexity will be moderate.*

---

*With the LPC analysis frame rate of 50 frames/s, one needs about 250 bits/s to transmit the prediction error energy.

The logic rule for variation of the multiplication factor β_n is shown in Table 6. This rule is essentially the same as the rule originally used by Winkler in his high information delta modulation.[31]* To reduce the slope overload noise in tracking the input signal by ADM, particularly when the low-passed residual increases or decreases quickly, a modification of the logic rule was made: Whenever the sign e_n changes after four consecutive signs of the same polarity, the step size Δ_n remains the same as Δ_{n-1} rather than being reduced to $0.66\ \Delta_{n-1}$. Otherwise, the logic rule is determined on the basis of three consecutive sign bits, as in Table 6.

The increase-step multiplication factor I = 1.5 and the decrease-step multiplication factor D = 0.66 were chosen because with these values the ADM seemed to track the input signal well, and the quality of the synthesized speech was the best. In choosing the increasing and decreasing factors, I and D, one must impose the following condition to ensure the stability of the decoded waveform:

$$ID \leq 1 \quad . \tag{25}$$

Note that, when I = D = 1, the hybrid companding of the ADM system becomes syllabic. Also, with I = D = 1 and the basic step size, Δ_o, constant at all times, the companding of the delta modulator becomes linear. The performance of the ADM system with different compandings is being investigated, with the SNR as the performance criterion.

In varying the ADM step size by the logic rule, the maximum and minimum have been set to certain values, typically, $\gamma_{n,max} = 9$ and $\gamma_{n,min} = 1$, respectively. Limitation of the maximum step size is required to prevent unnecessarily large overshoots when the input signal

---

*However, our ADM has different multiplication factors.

Table 6

ADM LOGIC RULE

e_n	e_{n-1}	e_{n-2}	e_{n-3}	e_{n-4}	Multiplication Factor (θ_n)
+	+	+			1.5
-	-	-			1.5
-	-	+			1
+	+	-			1
-	+	+			0.66
+	-	-			0.66
-	+	-			0.66
+	-	+			0.66
-	+	+	+	+	1
+	-	-	-	-	1

becomes constant after abrupt change. Consequently, a stable condition is achieved quickly. However, if the maximum step size were set too low, an excessive slope overload noise might result in tracking the input signal with the steep slope. Hence, one must consider the trade-offs in setting the maximum value of the quantizer step size. On the other hand, the purpose of setting the minimum step size is to prevent the "dead zone" effect. If the minimum step size were set too high, an excessive granular noise might result, while if the minimum step size were too low, the response to the sudden change of the ADM input might be too slow.

The performance of ADM in encoding the low-passed residual with cutoff frequency of 400 Hz was, as expected, better than with the cutoff frequency of 800 Hz, in terms of the signal-to-noise ratio. The rms SNR of ADM for the 400-Hz signal was 6 dB better than that for the

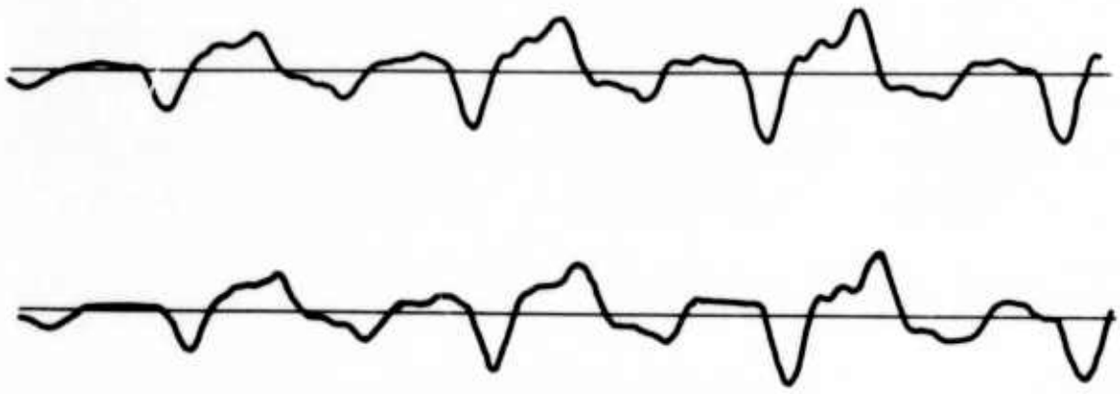
800-Hz signal at the sampling rate of 5 kHz. A detailed analysis of ADM performance is being made. Figure 25 shows the ADM input signals band-limited to 400 Hz and 800 Hz and their ADM-decoded waveforms after the low-pass smoothing filtering. The ADM sampling rate was 6.8 kHz.

One may observe from Figure 25 that an ADM waveform peak generally occurs about five samples after the corresponding peak of the input waveform.* This delaying effect is attributable to inherent properties of DM and of low-pass filters; i.e., a one-bit delay always occurs in ADM because of the delayer in the feedback branch of the ADM main loop, as can be seen in Figure 24. Additional delay occurs from the filtering effect of the low-pass smoothing filter. To compensate this delay effect, the LPC coefficients were delayed at the synthesizer for the same amount as the excitation signal is delayed due to low-pass filtering and ADM.

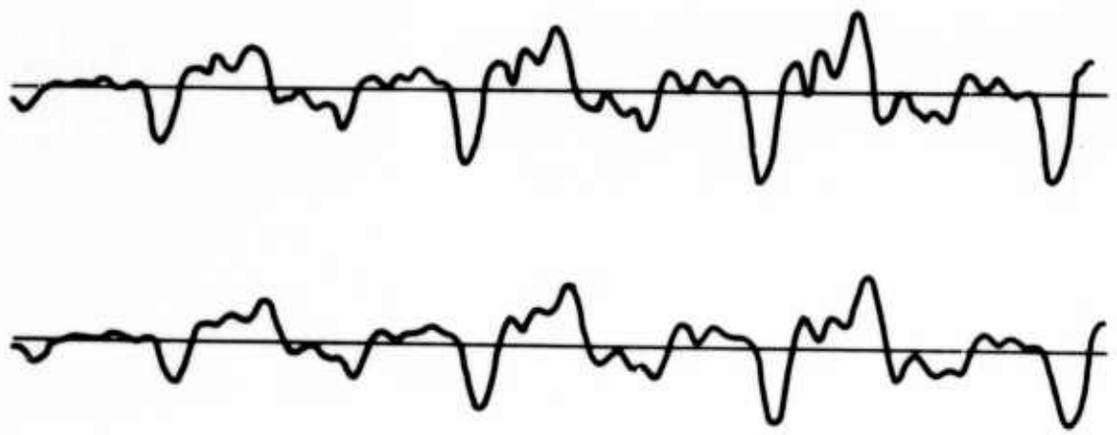
One drawback of a delta modulator is its transient effect at the beginning of encoding; i.e., the waveform of a delta modulator starts, in general, at an arbitrary amplitude level and "catches up" with the input signal only after a finite time. We have observed this effect in the ADM waveforms, but the transient time has seemed minimal and no degradation of the synthesized speech due to the effect has been detected. The hybrid companding of our delta modulator in this case should again work better than any other companding. Since the basic step size of the ADM quantizer is set at the syllabic rate according to the average energy of the input signal, the transient time with hybrid companding is, on the average, shorter than with either instantaneous or syllabic companding.

---

*The magnitude of the delay depends on the sampling frequency and the low-pass filter cutoff frequency.



(a) INPUT BAND-LIMITED TO 400 Hz



(b) INPUT BAND-LIMITED TO 800 Hz

SA-1526-46

NOTE: Sampling frequency 6.8 kHz.

FIGURE 25 COMPARISON OF ADM INPUT (TOP) AND DECODED (BOTTOM) WAVEFORMS

The ADM-decoded output has been smoothed by a low-pass filter with 800-Hz cutoff frequency. The process of smoothing removes much of the overshoots and granularities of an ADM wave and consequently improves the quality of the synthesized speech significantly.

e. Interpolator and Up-Sampler

If the residual has been down sampled before ADM coding, we must restore the original sampling frequency before feeding the residual into the LPC synthesizer. To up-sample we have used a linear interpolation; i.e., for the samples x_n and x_{n+1} we have generated a new sample, x_k, by

$$x_k = \frac{x_n + x_{n+1}}{2} \quad , \quad n < k < n + 1 \quad . \tag{26}$$

f. Spectral Flattener

In our simulation of the RELP system, the highest-frequency component of the ADM-decoded residual at the receiver was assumed to be 800 Hz for a telephone-line speech signal and 400 Hz for an unfiltered speech signal. Therefore, it is clear that the higher-frequency harmonics, at least up to 4 kHz of the residual, must be recovered before the residual is used as the excitation signal to the LPC synthesizer. It is well known that the higher-order harmonics of a signal may be generated through a nonlinear distortion process by feeding the signal to an instantaneous nonlinear device with zero memory, such as a ν_{th}-law device.[34] In our simulation we have used two types of nonlinear devices for spectral flattening: an asymmetrical linear full-wave rectifier and a harmonic generator using the concept of wideband frequency modulation (FM).

1) Asymmetrical Linear Full-Wave Rectifier

A block diagram of the spectral flattener with an asymmetrical linear full-wave rectifier is shown in Figure 26. For an input signal, x_n, the output, $\hat{x}_n$, of the rectifier is given by

$$\hat{x}_n = 0.8 \mid x_n \mid \quad , \quad \text{if } x_n \geq 0 \quad , \tag{27}$$

$$\hat{x}_n = 0.2 \mid x_n \mid \quad , \quad \text{if } x_n < 0 \quad . \tag{28}$$

The rectified residual is then differenced twice to enhance the high-frequency components,

$$\bar{x}_n = \hat{x}_n - 2G\,\hat{x}_{n-1} + G^2 x_{n-2} \quad , \tag{29}$$

or in Z domain

$$\bar{X}(z) = (1 - G z^{-1})^2 \hat{X}(z) \quad , \tag{30}$$

where G is a gain factor.

Note that previously we used a half-wave linear rectifier and a double differencer as a spectral flattener. In this case we observed that the synthetic speech wave occasionally had null regions in unvoiced portions while the original wave did not. We determined that this was caused by the effect of half-wave rectification. The reason is that in the case of using a half-wave rectifier as a spectral flattener the excitation signal of the LPC synthesizer is a positive half-wave of the low-passed residual and sometimes has unusually long null periods in an unvoiced sound, such as in /s/ of "grasp." Consequently, no excitation occurs in this null period. To remedy this problem the rectifier has been modified to include also some negative portions of the residual signal, as shown above.

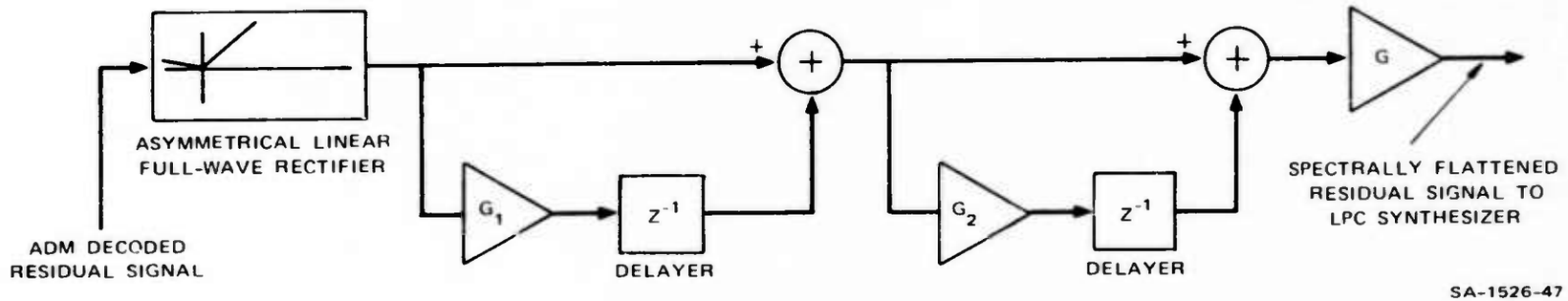


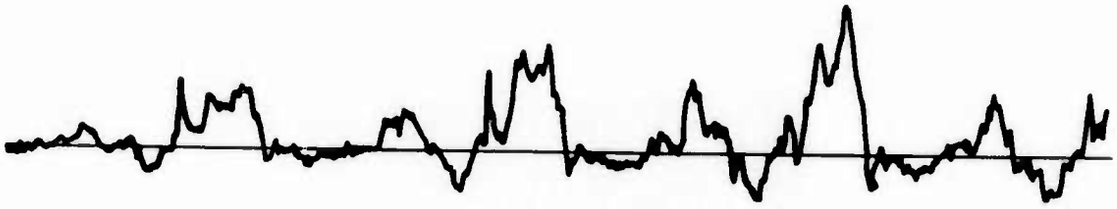
FIGURE 26 SPECTRAL FLATTENER—ASYMMETRICAL LINEAR FULL-WAVE RECTIFIER AND DOUBLE DIFFERENCER

The outputs of the asymmetrical linear full-wave rectifier for the phonemes /o/ in "oak" and /s/ in "strong" are shown in Figure 27. Shown in Figure 28 are the spectra of the original residual and the spectrally flattened low-passed residual, together with the time waveforms for the phoneme /i/ in "Pete."* When the spectrum of the original residual and the spectrally flattened spectrum are compared, the latter seems whiter, which is desirable for better quality of speech. Also, note that although the temporal waveforms of the two residuals look entirely different, the spectra are very similar. Hence, it is clear that the reason why we are getting a good quality of synthesized speech is that the residual encoding in the RELP system is essentially a spectral-matching process. This is in contrast to other residual-encoding methods, such as DPCM without spectral flattening, which attempts to match the waveform. One might note that linear prediction with an all-pole filter can be interpreted as a spectral-matching process. Therefore, it may be concluded from the observation of the waveforms that what is important in LPC synthesis is the frequency content of the synthesizer input signal, but not the signal waveform itself. In other words, the LPC synthesizer will do a good job of spectral matching as long as all the correct frequency harmonics are present in the excitation signal, i.e., the system is waveshape independent.

Although the asymmetrical linear full-wave rectifier with a differencer is simple and easy to implement, it is not the best spectral flattener for generating high-frequency harmonics. For this reason we have also considered the possibility of using the FM theory for a better spectral flattener. This and other possible methods of harmonic generation are discussed next.

---

*The rectifier used in this case was a linear half-wave rectifier.



(a) /o/ IN "OAK"

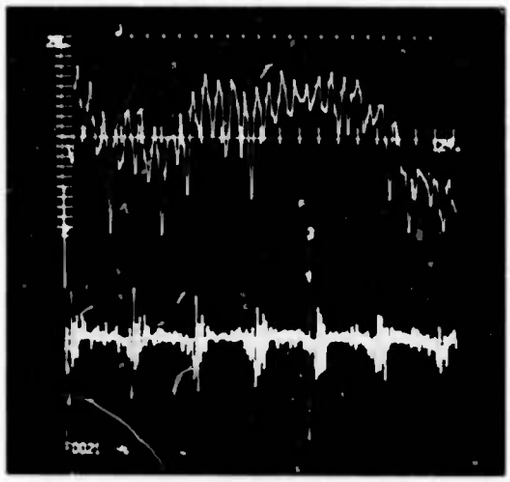
(b) /s/ IN "STRONG"

SA-1526-48

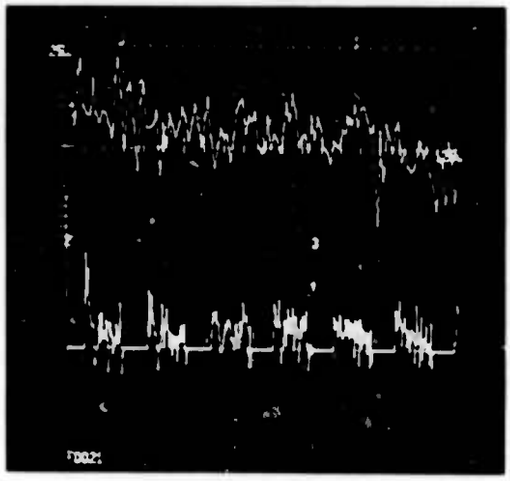
FIGURE 27 SPECTRALLY FLATTENED RESIDUAL WAVEFORMS

2) Frequency Modulator Spectral Flattener

Harmonics of the fundamental pitch can be created by using various nonlinear operators. Achievement of the desired harmonic level can represent a serious problem. For example, the glottal wave-shapes illustrated in Figure 29 do not have their higher harmonic content enhanced by a half-wave linear rectifier. The reason is simple: No



(a) ORIGINAL RESIDUAL



SA-1526-49

(b) SPECTRALLY FLATTENED LOW-PASSED RESIDUAL (LINEAR HALF-WAVE RECTIFIER USED)

FIGURE 28 LPC RESIDUAL SPECTRA (TOP) AND TEMPORAL WAVES (BOTTOM) OF /i/ IN "PETE"

FIGURE 29 LOW-PASS-FILTERED RESIDUAL WAVEFORM, WHOSE HARMONIC CONTENT CANNOT BE ENHANCED BY HALF-WAVE LINEAR RECTIFICATION

sharp waveform "corners" or discontinuities are created by the nonlinearity and, thus, no higher harmonics are generated.*

It is well known in communication theory that frequency modulation is an extremely effective and controllable method of generating harmonics. Figure 30 illustrates the block diagram of a spectral flattener based on the FM principle. Note that this system corresponds to FM with a zero carrier frequency (the left-hand or harmonic path). The baseband signal from the adaptive delta demodulator output is passed to the output in undistorted form. However, the harmonic generation path includes the form $y = \cos(\beta X + \pi/4)$, where the term $\pi/4$ has been included to guarantee the presence of both even and odd harmonics (see Figure 31). From FM theory recall that

$$\cos(\beta \sin wt) = J_0(\beta) + 2J_2(\beta) \cos 2wt + 2J_4(\beta) \cos 4wt + \ldots \quad (31)$$

and

$$\sin(\beta \sin wt) = 2J_1(\beta) \sin wt + 2J_3(\beta) \sin 3wt + \ldots \quad , \quad (32)$$

---

*The reader should recall from his first encounter with Fourier series that harmonics of the fundamental are required to fill in the corners of a square wave. Thus, it is the sharp corners that create harmonics.

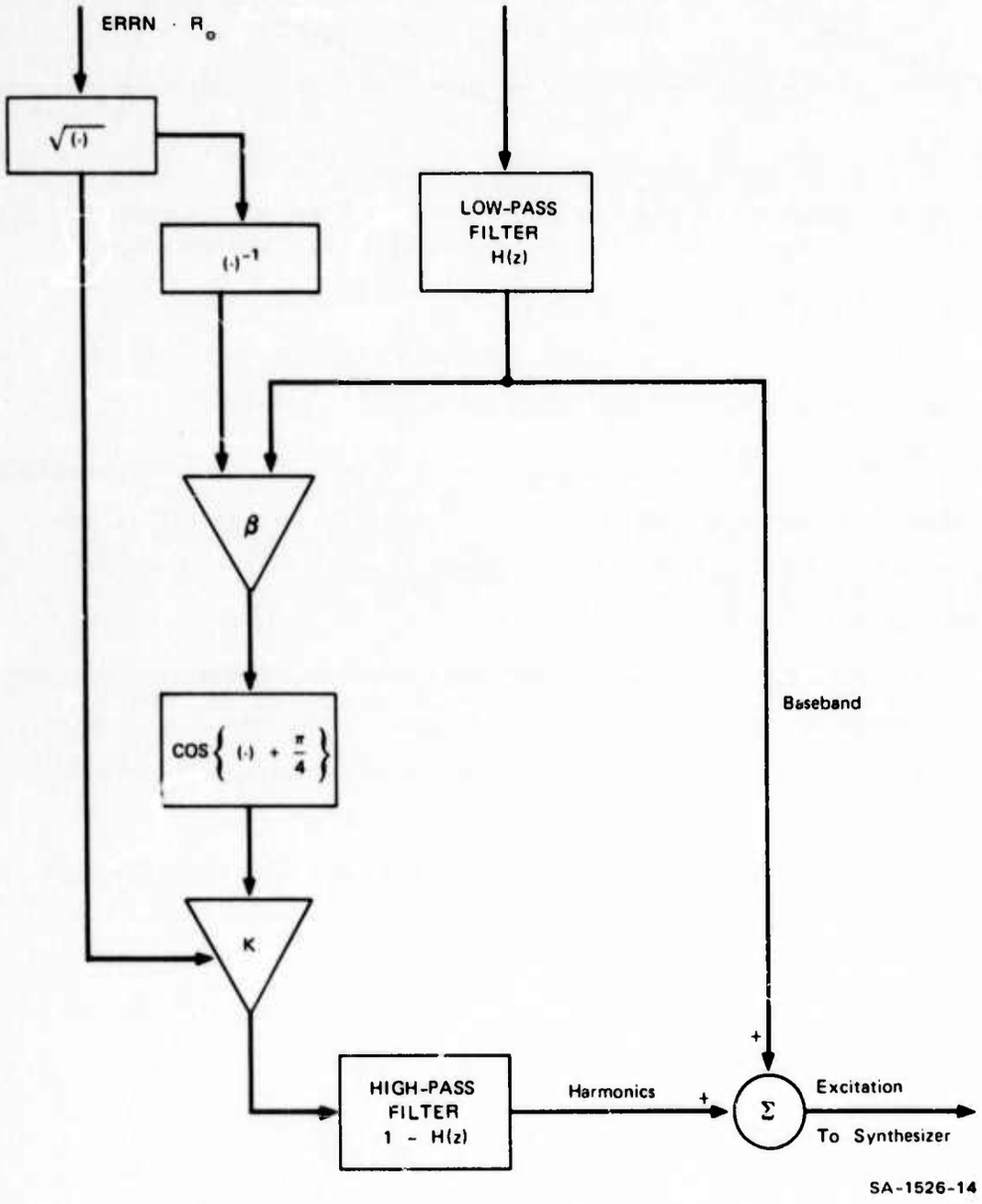


FIGURE 30 FM HARMONIC GENERATOR SPECTRAL FLATTENER

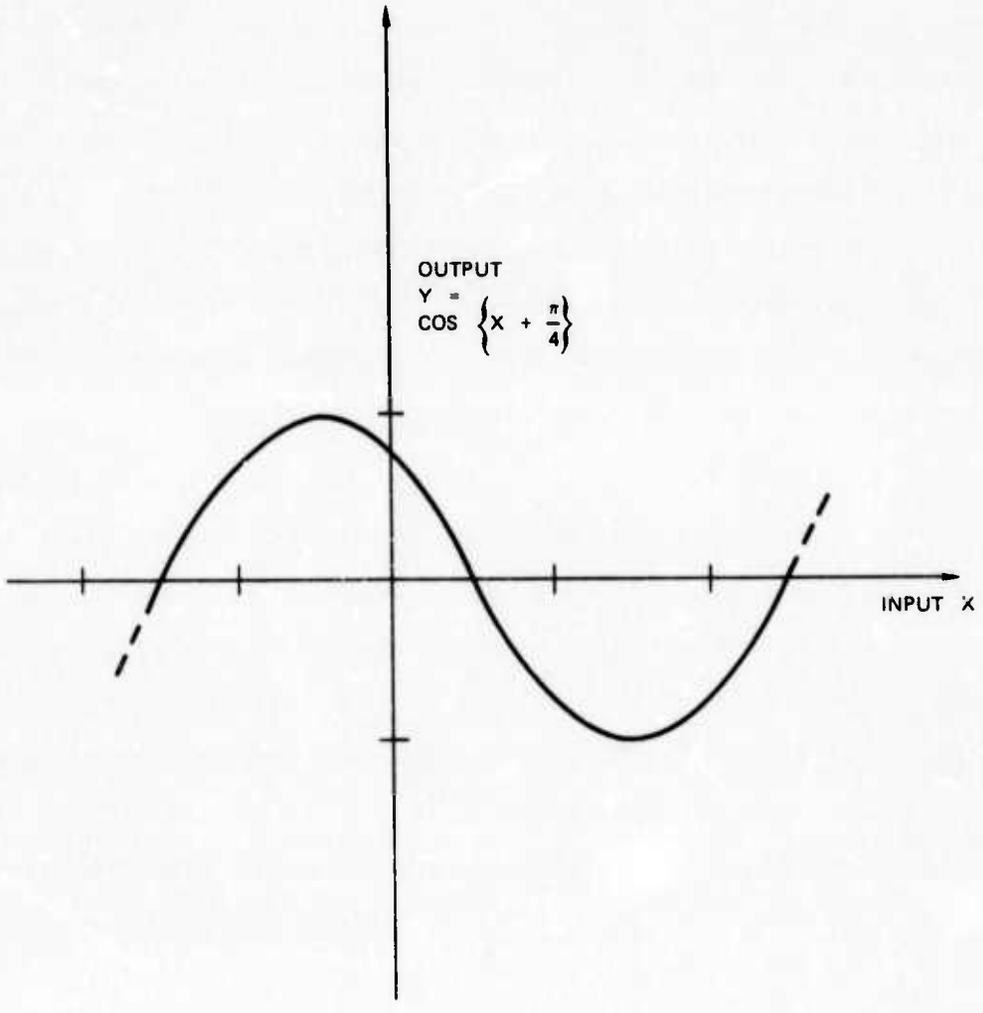


FIGURE 31 INPUT-OUTPUT DIAGRAM OF THE ZERO-MEMORY FM HARMONIC GENERATION NONLINEARITY

where $J_n(\beta)$ is the Bessel function of the first kind. Thus, a phase shift of 45 degrees will ensure the presence of both even and odd harmonics in the output. Figure 31 illustrates the FM harmonic generator zero-memory nonlinearity.

Note that the output of the FM nonlinearity is high-pass filtered so that no nonlinear distortion appears in the baseband at the output. The gain factor, β, of the FM nonlinearity is inversely

proportional to $\sqrt{ERRN \cdot R_o}$, the root-mean-square energy in the output from the adaptive delta demodulator. Thus, an automatic gain eontrol (AGC) action is used, which results in a relatively fixed harmonie strueture independent of the power of the input signal. The FM nonlinearity produces constant output power independent of the input level. Thus, it is necessary to follow this nonlinearity with a variable-gain amplifier, K, proportional to $\sqrt{ERRN \cdot R_o}$. Use of this amplifier permits the harmonic level to appropriately track the level of the input, as does the baseband (right-hand path in Figure 30) channel.

Note that the FM nonlinearity does not sharpen the fall times illustrated in Figure 29. If thi: could be done, ideal performance would be achieved. The harmonic generator that would aecomplish this task would produce a sharp negative pulse at the fall time eorresponding to the glottal stop. Thus, it would be desirable to phase the harmonics so that they produce a pulse waveform. The FM harmonie generator is unable to produce this phasing. Therefore, some slight quality degradation may result. The high-frequency components encounter phase distortion such as may occur in a room with acoustie refleetions. Thus, the quality degradation should be minimal, corresponding to that in normal listening environments. In fact, the difference between ideal and FM harmonic distortion should be discernible only when listening with head phones.

To date, limited success has been achieved with the FM harmonic generator approach. High-frequency components are ereated and enhanced as predicted. However, the synthesized speech tends to have a gargling quality. It is hypothesized that this problem is eaused by the time distribution of the harmonics. It appears that they are bunched in time, producing a multiple exeitation phenomenon that is perceived as a slight gargle.

3) Alternative Spectral Flatteners

This section lists some alternative forms for spectral flatteners. It may be possible to improve quality by using one of these approaches. However, recent effort has been concentrated on the two techniques, the asymmetrical full-wave linear rectifier and the FM harmonic generator, already described.

First, the bandpass limiter bank technique has frequently been employed in VEVs. The desired harmonics are created by a nonlinearity. A constant envelope spectrum is then produced through use of a bank of bandpass limiters operating on the output of the nonlinearity. This approach is guaranteed to produce a signal rich in high-frequency harmonics. However, it requires system complexity that should be avoided, if possible. The output of the bandpass filter bank has a fixed level, so it is necessary to apply an adjustable gain to this signal.

Second, one might consider employing a center-clipping nonlinearity, as illustrated in Figure 32. Examination of this nonlinearity shows that it certainly will create sharp corners and enhance rise times. This concept also makes good intuitive sense, since center-clipping is known to destroy formant information. Thus, this nonlinearity should tend to enhance the desired pitch information. Problems associated with the approach are:

- Correct choice of threshold.
- Finding a method of establishing the correct power for driving the synthesizing filter. Note that the proposed design for the center-clipping nonlinearity is slightly different from the conventional center clipper (see Figure 32).

Third, one could consider employing a nonlinear phase-versus-frequency all-pass filter to distort the glottal waveshape of

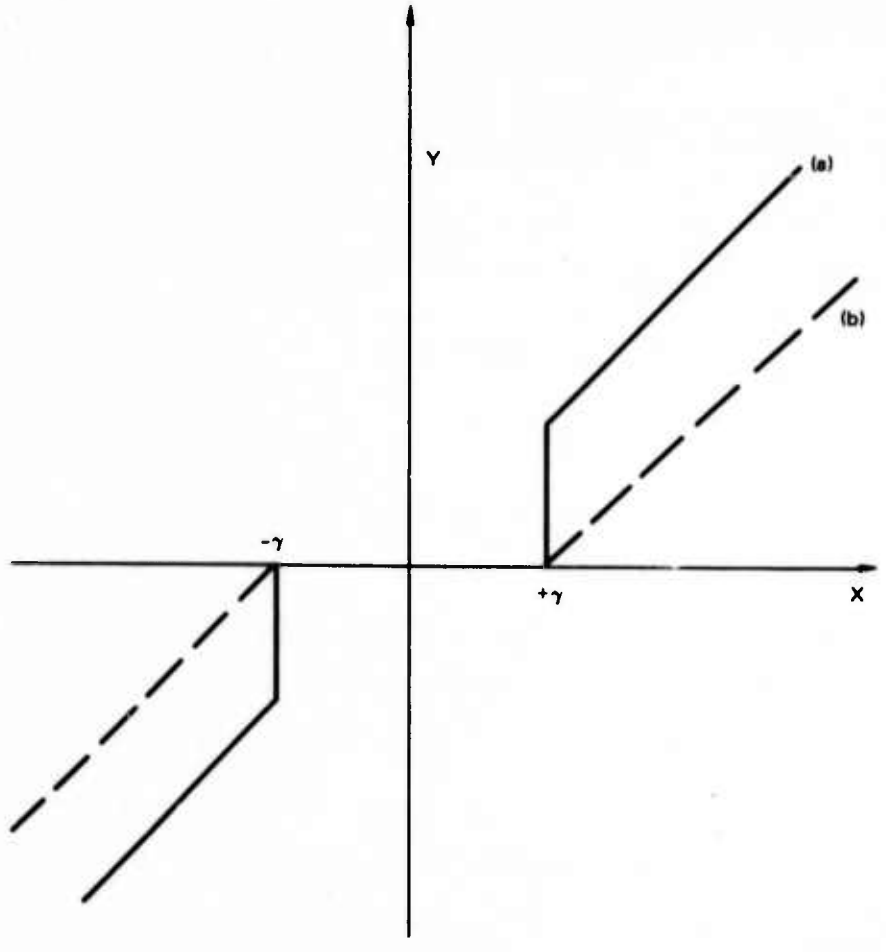


FIGURE 32 INPUT/OUTPUT CHARACTERISTIC FOR CENTER CLIPPERS (a) PROPOSED FORM, AND (b) CONVENTIONAL FORM

Figure 29 in such a manner that a half-wave linear rectifier would enhance the harmonic content. This approach is reasonably simple and should offer some modest improvement for waveforms similar to that of Figure 29. However, note that under many circumstances the acoustic environment automatically provides the desired phase distortion. In these cases this proposed design offers no improvement.

Fourth, if glottal-stop waveforms similar to the waveform of Figure 29 could be consistently generated, it would be possible to use logical slope enhancement. With this approach a pattern recognition circuit would recognize the fall time (glottal-stop) waveform and replace it with a sharper cutoff waveform.

The glottal stop could be recognized by a number of large positive values, followed by a few descending values, and then a number of low values. Such a pattern recognition system could not be expected to work perfectly. However, for a large percentage of the time the harmonic content could be significantly increased.

Finally, a number of nonlinearities could be tried. Bogner and Hashed reported on a linear harmonic generation technique that appears particularly desirable since it avoids the power control problem.[35] That is, the harmonic levels scale linearly with the input level. This approach for selecting an optimum nonlinearity will be pursued in conjunction with the above techniques that require a zero-memory nonlinearity.

g. LPC Synthesis

Since a residual signal (rather than pitch pulses) is used for excitation of the synthesizer, LPC synthesis is done pitch asynchronously in the RELP system. The synthesizer is mathematically the inverse of the prediction filter, A(z), and may be implemented in several ways. For instance, if a synthesizer that is the direct inverse form of the prediction filter, $1/A(z)$, is desired, the received reflection coefficients* are transformed recursively into the prediction filter coefficients, and

---

*It is to be understood that the coefficients received from the RELP transmitter are the reflection coefficients rather than the predictive coefficients.

then the synthesizing filter is formed. Also, it is possible to implement the synthesizer without transformation by using the reflection coefficients.[36] The synthesizing filter in this case becomes a lattice or ladder. We have chosen the ladder form of the synthesizer in the RELP system because it is known to yield greater numerical accuracy and a less complex sequence of arithmetic operation and also to give a simple stability check. The synthesizer of the RELP system is shown in Figure 33, where it is illustrated in lattice rather than ladder form.

The input signal to the synthesizer is, as stated previously, a spectrally flattened low-passed residual mixed with random white noise generated from a local random noise generator. The formula for adding random noise to the excitation signal is as follows:

$$\gamma_n = \bar{X}_n (1 - R) + \left(\frac{E}{N}\right)^{1/2} \cdot R \cdot \Lambda \qquad (33)$$

with

$$R = \frac{ERRN}{\ell} \quad , \quad 0.05 \leq R \leq 1 \quad , \qquad (34)$$

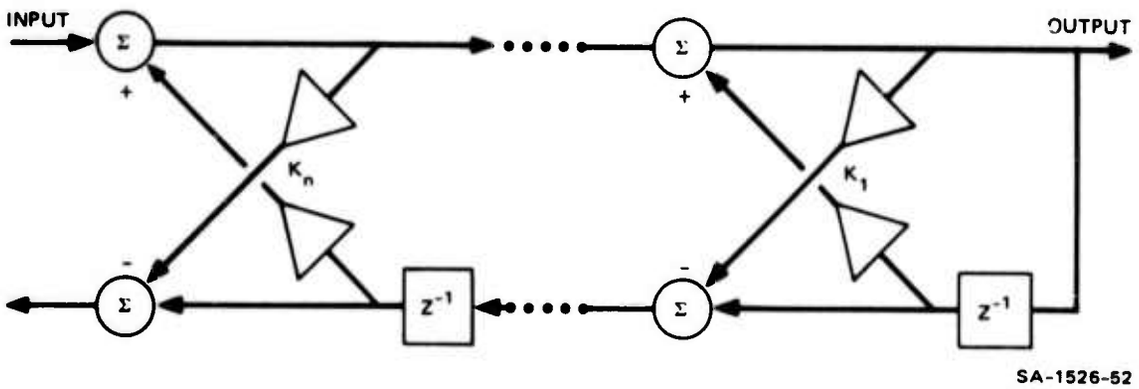


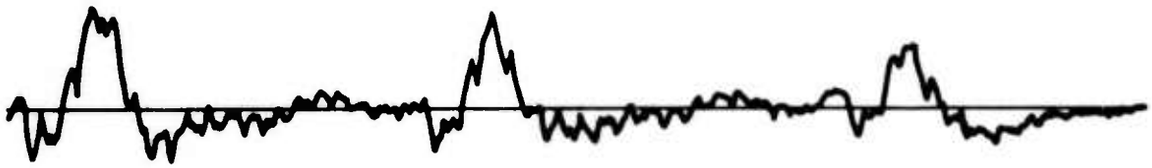
FIGURE 33 LPC LATTICE SYNTHESIZER

where χ_n denotes the synthesizer excitation signal with random noise; $\bar{X}_n$, the spectrally flattened residual signal; E, the prediction-error energy; N, the number of samples in one analysis frame; Λ, the output of the random noise (or number) generator; and ERRN, the normalized prediction-error energy. Note that the amount of random noise energy is controlled by the scale factor k in R and also by variation of the lower bound of R. It was determined that $k = 6$ and R = 0.1 yields the best quality of synthetic speech. With k less than 3, one can hear the effect of excessive noise in the synthesized speech.

Fujimura reported that even voiced signals have unvoiced (i.e., aperiodic) portions in the high-frequency range above 1000 Hz.[37] He further claimed that addition of a proper amount of unvoiced signal or random noise to the excitation signal definitely improves speech quality. Our experiment confirmed his latter claim. In an earlier phase of development of the RELP vocoder, the synthetic speech lacked in general the high-frequency energy--particularly in fricatives. However, as a result of adding random noise, the synthetic speech quality has been improved. This improvement can be seen in Figure 34, which shows the synthetic waveforms with and without random noise mixed with the excitation signal, along with an original waveform. It should be noted that adding aperiodic random noise does not completely correct the lack of energies of periodic high-frequency harmonics in the voiced signal, although it alleviates the problem to some extent. Of course, for the unvoiced signal the problem of lack of energy could be completely solved by adding random noise.

To feed a correct amount of the excitation energy to the LPC synthesizer, the magnitude of the excitation signal is controlled by the prediction error energy as

$$\bar{\chi}_n = \left(\frac{E}{W}\right)^{1/2} \chi_n \qquad (35)$$



(a) ORIGINAL

(b) SYNTHETIC WITHOUT RANDOM NOISE IN THE EXCITATION SIGNAL

SA-1526-53

(c) SYNTHETIC WITH RANDOM NOISE

FIGURE 34 COMPARISON OF SYNTHETIC WAVEFORMS OF /z/ IN "IS" WITH AND WITHOUT RANDOM NOISE MIXED WITH THE EXCITATION SIGNAL (SPEECH INPUT TO LPC ANALYZER WAS PREEMPHASIZED)

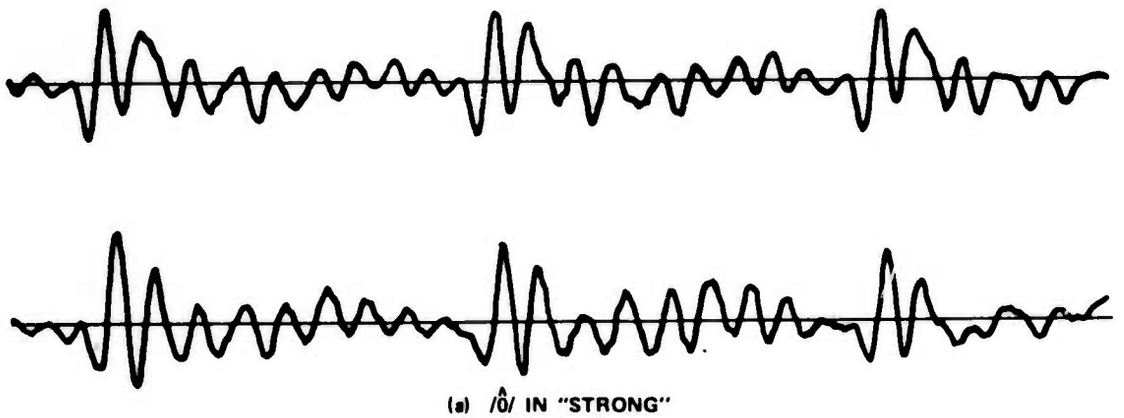
with

$$W = \sum_{n=0}^{N-1} \chi_n^2 \tag{36}$$

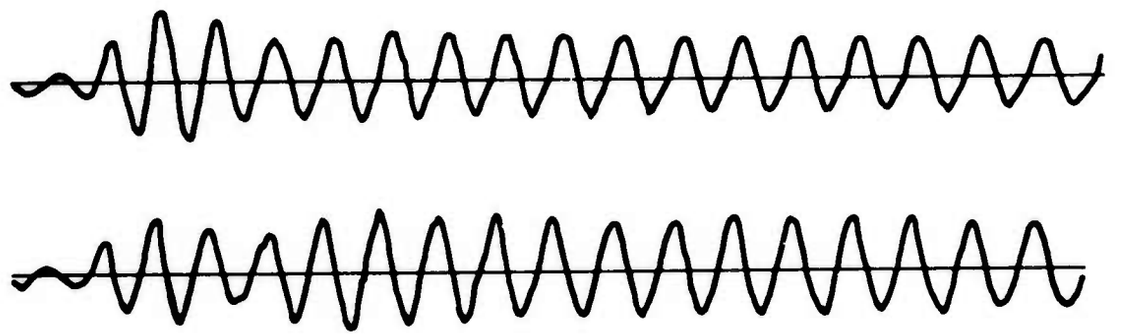
where $\bar{\chi}_n$ is the excitation signal with the gain control, E is the prediction error energy, and χ_n is the spectrally flattened residual mixed with random noise. Note that E and W are computed in each LPC analysis and synthesis block and that the prediction error energy, E, is also used in a syllabic companding of ADM coding of the residual signal. The gain control is actually not necessary most of the time. It is particularly effective, however, when occasional overshoots of the ADM wave cause click noises in the synthetic speech. In such a situation the gain suppresses the overshoots and thus no click noises occur.

We have used two kinds of input speeches in our simulation. One was generated without any background noise in a room with an ideal acoustic condition; the other was recorded simultaneously with a background utterance. The latter input speech was recorded in a room where acoustic condition was not ideal. The purpose of using the two different input speeches was to demonstrate that the RELP vocoder can be operated in any environment.

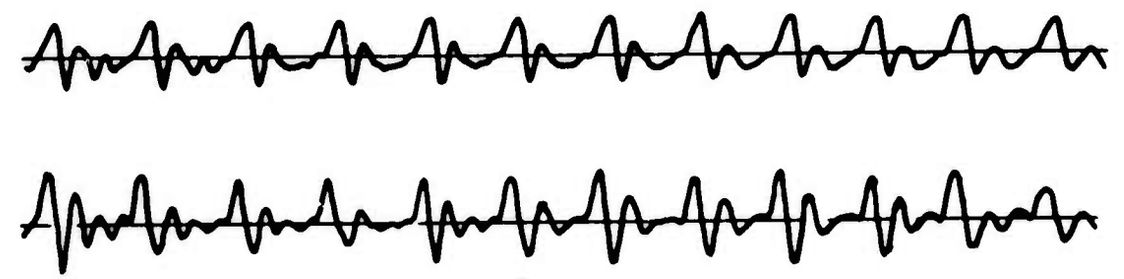
Figure 35 shows the original and synthetic speech waveforms of various phonemes. The original speech was recorded in a room with an ideal acoustic condition. The synthetic waveforms have been generated with the parameters summarized at the end of this section. Figure 36 shows the original and the synthetic speech waveforms of /p/ in "product," with /z/ in "dogs" superimposed. Because the input speeches were recorded deliberately in a room with a nonideal acoustic condition, some phase distortion can be seen in the original waveform. Even if two utterances are superimposed or are present simultaneously, the original and



(a) /$\hat{0}$/ IN "STRONG"



(b) /bi/ IN "BEGAN"

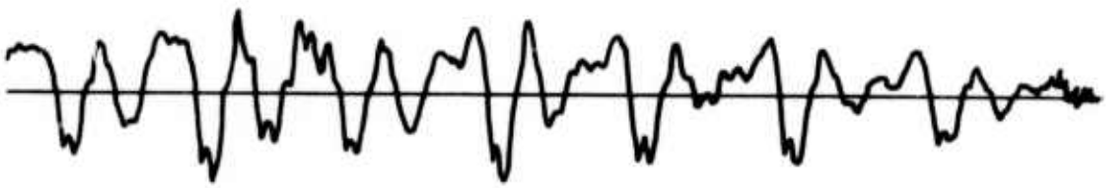


(c) /$\bar{1}$/ IN "WHILE"

SA-1526-54

FIGURE 35 COMPARISON OF ORIGINAL (TOP) AND SYNTHETIC (BOTTOM) SPEECH WAVEFORMS

(a) ORIGINAL



(b) SYNTHETIC

SA-1526-50

FIGURE 36 COMPARISON OF ORIGINAL AND SYNTHETIC WAVEFORMS OF /p/ IN "PRODUCT" WITH BACKGROUND SPEECH OF /z/ IN "DOGS" SUPERIMPOSED

the synthetic waveform clearly look alike. One interesting observation from the experiment with multiple speeches was that, when the amplitude of the background utterance was low, the LPC synthesizer suppressed it; thus one could not hear the background utterance in the synthetic speech, even though both utterances could be heard in the original. Such a capture effect can be extremely useful.

In Figures 35 and 36 one can see that the synthetic speech waveforms slightly lack high-frequency components compared with the original waveforms. This effect results from heavy low-pass filtering of the

residual signal. Previously, we did not preemphasize the input speech before LPC analysis and used the spectrally flattened residual (without random noise mixed) as the excitation signal of the LPC synthesizer. In that case, the lack of high-frequency energy in the synthetic speech was a serious problem. As a result of preemphasis and mixing random noise, however, the problem has been largely overcome. The effect of the slight lack of high-frequency components that still exist in the synthetic speech is hardly perceptible in most of the cases. Figure 37 shows the synthetic waveforms with and without preemphasis, along with the original waveform. We have not deemphasized or integrated the synthetic speech output for the obvious reason that the inverse operation of preemphasis attenuates the high-frequency energy of the synthetic speech.

The recorded audio tape of the original and synthetic utterances generated by the simulated RELP vocoder accompanies this report. Table 7 summarizes the important parameters and specific methods used in the RELP simulation.

h. Computer Flow Chart

The RELP vocoder system has been simulated on an Interdata 70 minicomputer. The configuration of the machine is as follows:

- 48K bytes of memory
- One Disc-Diablo with 2.5 megabytes
- One tape drive
- One graphics terminal: Tektronix 4010 with hard copy
- One teletype
- One custom-built 12-bit A/D and D/A converter.

The computer program was written in FORTRAN with double precision floating-point arithmetic. Figure 38 is a flow chart showing the general flow of the program for the system. Figure 39 is a flow

(a) ORIGINAL

(b) SYNTHETIC WAVE WITHOUT PREEMPHASIS

(c) SYNTHETIC WAVE WITH PREEMPHASIS

SA-1526-51

FIGURE 37 COMPARISON OF SYNTHETIC WAVEFORMS OF /I/ IN "IS" WITH AND WITHOUT PREEMPHASIS OF SPEECH INPUT TO LPC ANALYZER

Table 7

SUMMARY OF PARAMETERS AND METHODS

Transmission rate*	Residual	5,000 bits/s†
	Coefficients	2,750 bits/s
	Gain	250 bits/s
	Total	8,000 bits/s
Input bandwidth	3.2 or 4 kHz	
Input sampling rate	6.8 or 10 kHz	
Window	Hamming window	
LPC analysis	Autocorrelation method	
Analysis frame rate	50 frames/s	
Number of LPC coefficients	Ten for 3.2-kHz speech input and 12 for 4-kHz input	
Cutoff frequencies of low-passed residual	400 Hz and 800 Hz	
Coefficients and gain encoder	Pulse code modulation	
Residual encoder	Adaptive delta modulation with hybrid companding	
ADM sampling rate	3.4 to 6.8 kHz	
Spectral flattener	Asymmetrical linear full-wave rectifier	
LPC synthesizer	Itakura's lattice filter	

*The transmission rate varies from 6K to 9.6K bits/s. The above calculation is typical for an input speech with the bandwidth of 4 kHz.

†First two coefficients are coded with six-bit quantization levels. The next three are coded with five-bit levels, and the remaining seven are coded with four-bit levels. Hence, with the analysis frame rate of 50 frames/s, we have $(6 \times 2 + 5 \times 3 + 4 \times 7) \times 50 = 2{,}750$ bits/s.

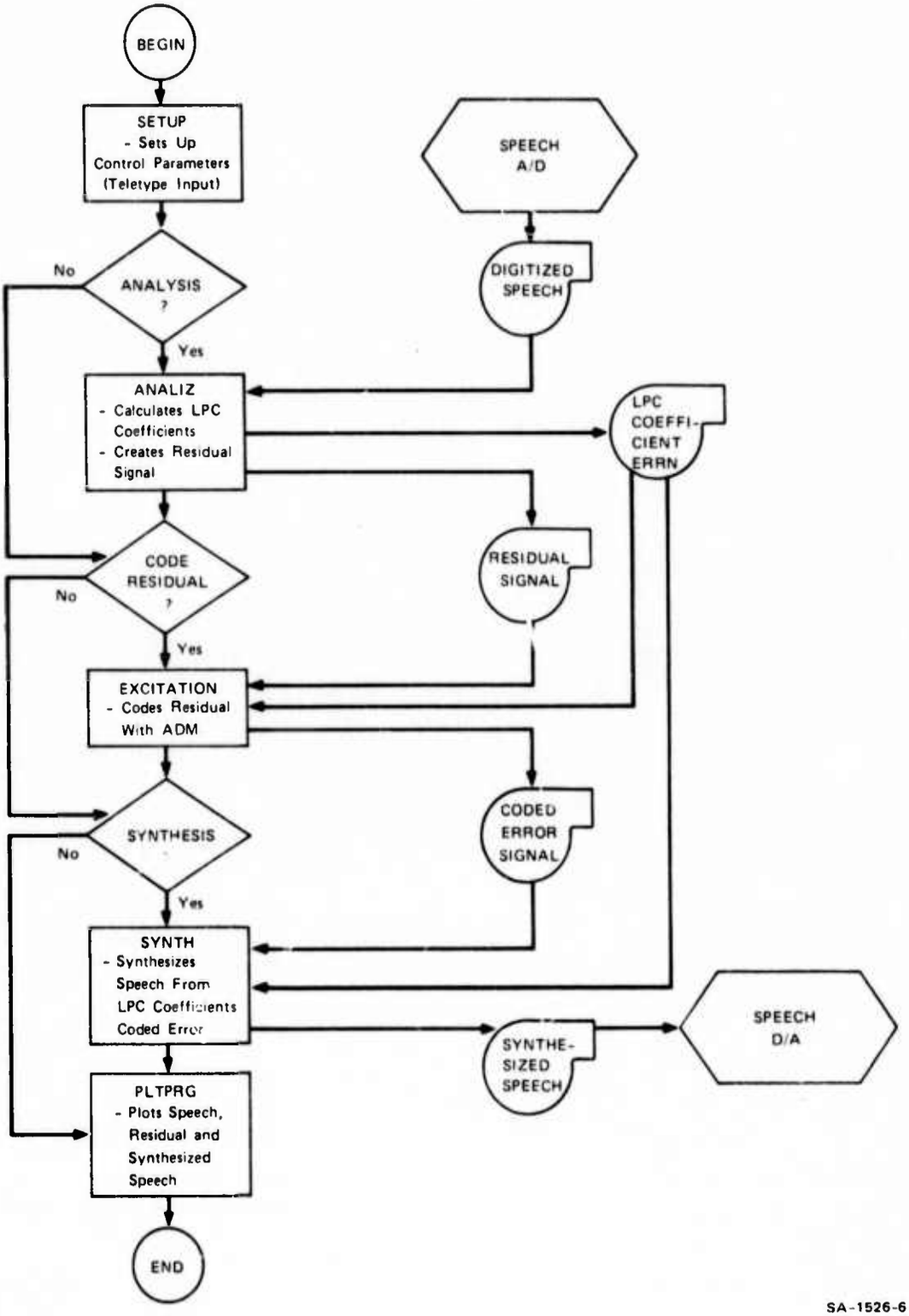


FIGURE 38 FLOW CHART OF THE COMPUTER PROGRAM FOR THE OVERALL RELP SYSTEM

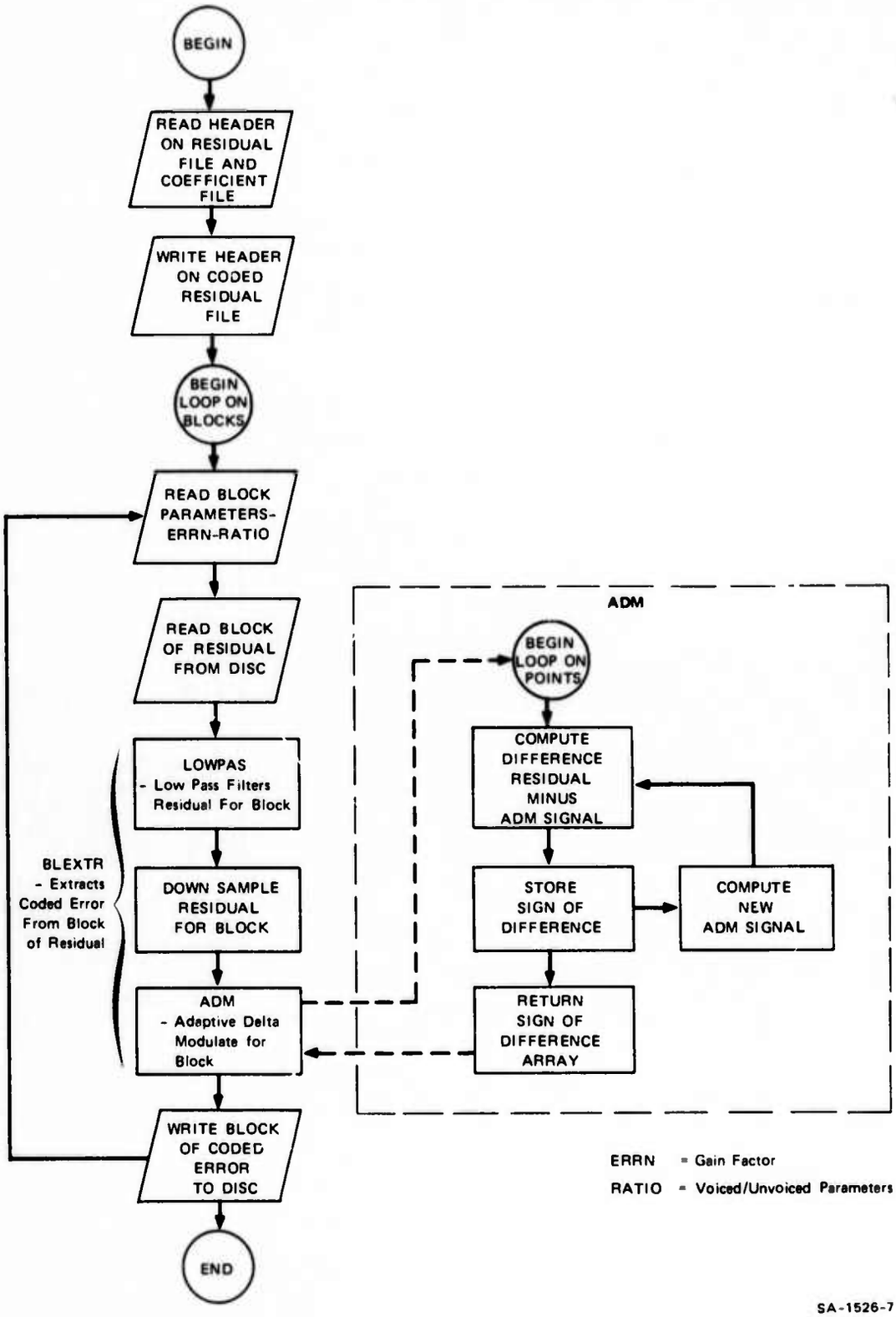


FIGURE 39 FLOW CHART SHOWING THE RESIDUAL ENCODING BY ADM AT THE TRANSMITTER

chart of the subroutine EXCITATION. This is the routine that codes the residual signal by ADM after low-pass filtering. Figure 40 shows a flow chart of the subroutine SYNTH. This is the routine that decodes the sign bits at the receiver to obtain the estimate of the input signal, spectrally flattens it, and then feeds it into the LPC synthesizer.

We have not made any significant effort to reduce the computer processing time, since the system is still in the developing stage. The program now runs much slower (approximately 100 times) than real time.

3. Discussion and Conclusion

We have demonstrated the capability of operating the RELP vocoder in any operating condition. Thus, the RELP system has the significant advantage of operating in a nonideal condition, as compared with a pitch-excited LPC. Of course, this is possible because no pitch extraction is necessary in the RELP system. A pitch-excited LPC has the advantage of saving the bandwidth by a factor one half* over the RELP vocoder, and its quality of synthesized speech suffers little degradation with accurate pitch markings. However, a disadvantage is that it requires pitch extraction; therefore, its performance could deteriorate unacceptably in a nonideal operating condition. For instance, pitch extraction is extremely difficult, if not impossible, in the presence of background noise or with multiple simultaneous speeches. Since the quality of any pitch-excited vocoder is highly dependent on the accuracy of the pitch information, in such situations the performance of a pitch-excited LPC would be unsatisfactory.

---

*We assume here that the pitch-excited LPC is operated between 3K and 4K bits/s.

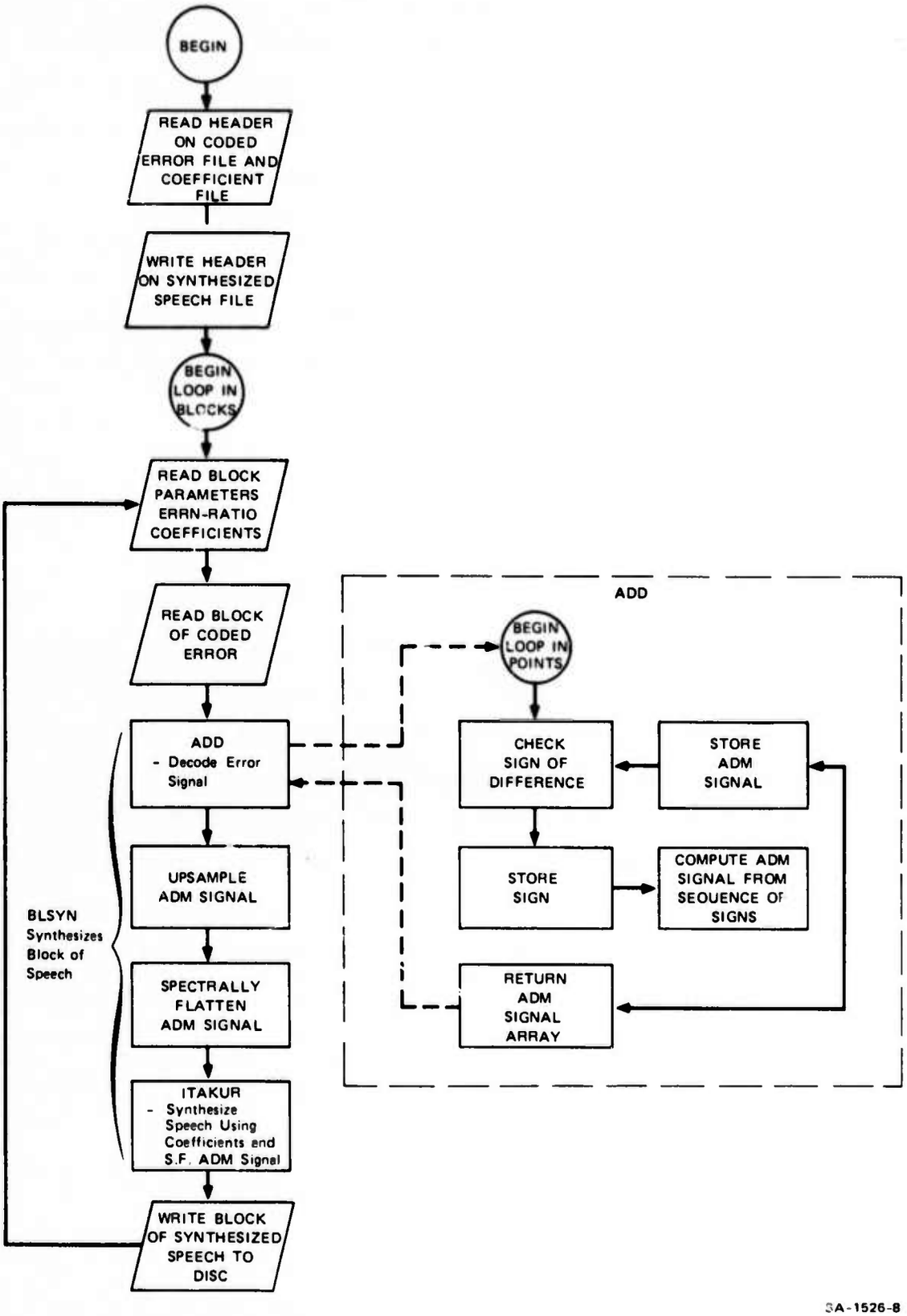


FIGURE 40 FLOW CHART SHOWING DECODING OF ADM SIGNAL, SPECTRAL FLATTENING, AND LPC SYNTHESIS AT THE RECEIVER

One may note, as seen in Figure 18, that each functional block of the RELP vocoder system is highly modularized, and the LPC analyzer and synthesizer are exactly the same as those of a pitch-pulse-excited LPC. Therefore, assuming that a good pitch extractor is available for incorporation in our system, our vocoder system can be operated in a hybrid mode; i.e., the LPC synthesizer can be excited by either pitch pulses or nonlinearly processed residual, depending on the quietness level of the system-operating environment. Thus, one can take advantage of both systems. If a fully reliable pitch extractor becomes available in the future, it can replace the residual encoder in the RELP system without excessive cost for hardware modification.

It should be noted that simulation has been done with the LPC analyzer and the synthesizer placed back to back. Therefore, no transmission errors of coefficients, residual, and gain have been assumed.

In conclusion, it is clear from demonstration of the recoded audio tape that the RELP algorithm we have developed offers much promise as an alternative to a pitch-excited linear predictive coding. It gives the lowest data rate among the residual-excited vocoders, while yielding synthesized speech of good quality.

IV CONCLUSIONS

The results of our research can be summarized as follows. The pitch-extraction problem is very difficult. It would appear that it is always possible to postulate an operating environment that will force a pitch extractor to make errors. In contrast, the residual-encoding approach is robust and tolerant of practical difficulties. From the viewpoint of a practicing communication engineer familiar with real world problems, the residual-encoding approach is much more sensible for the near future. This is true in spite of the fact that the residual-encoding approach may require as high as 9.6K bits/s compared with perhaps 2.4K bits/s for the pitch-extraction approach.

However, if the operating environment can be controlled (e.g., background noise eliminated by the use of high quality acoustic equipment, such as noise-canceling microphones), pitch extraction makes very good sense because of the lower data rate system that results.

Research in the long-term memory area indicated that pitch extraction based on delay-lock loop tracking of the glottal pulse was not very promising. Many problems, particularly those of acquisition, exist. In addition, it is difficult to guarantee the existence of a glottal pulse in the residual.

Our research demonstrated that by performing the right type of LPC analysis (use of spectral averaging) it was possible to recover the glottal waveshapes when the speech was recorded under good acoustic conditions, i.e., no phase distortion due to multipath. In these cases a simple peak-picking, time-domain pitch extractor may be adequate. However, the

Preceding page blank

discriminant approach based on the short-term residual energy is undoubtedly more reliable.

The formant-isolation approach to pitch extraction proved to be complex and sometimes unreliable, although the concept is attractive. However, the approach was very helpful as an experimental tool in hand marking of pitch pulses; it speeded this effort considerably. An extremely useful by-product of our formant-isolation research is a data base of very high quality pitch marks and resultant very high quality synthetic speech. The outputs of more practical real-time pitch extractors can be compared with this data to permit both analytical and subjective quality performance measures.

We halted research efforts in the long-term memory area to concentrate our resources on the short-term encoding approach. The result has been the development of the RELP system, which is the LPC equivalent of the voice-excited vocoder. The RELP system has been demonstrated to work successfully in a hostile environment, e.g., two simultaneous speakers. The quality of synthetic speech of the RELP vocoder is quite good at the transmission rate of 9.6K bits/s. Unlike other residual-excited coders, such as DPCM, the variation of the transmission rate with the RELP system is flexible and gradual. It is possible to have the transmission rate as low as 6K bits/s without significant degradation of quality. Since a major goal of this task was to develop an excitation encoder capable of operating in an imperfect environment, the decision to concentrate our efforts in the short-term memory encoding area has proved wise.

One important by-product of research on the RELP vocoder is the development of the ADM system with hybrid companding, which could compete with the best ADM system currently available, i.e., the continuously variable slope delta modulator (CVSDM). Although the ADM with hybrid

companding has been developed as an encoder of the residual signal, it should be noted that the system can be used as well for directly encoding a speech signal.

Appendix

BASIC PROGRAMS--CPMP5 AND CPMP6

Preceding page blank

Appendix

BASIC PROGRAMS--CPMP5 AND CPMP6

In this appendix we present the listings for two BASIC programs--CPMP5 and CPMP6--used to test the feasibility of extracting pitch by locating force-free (or excitationless) periods. Both programs used the non-Toeplitz analysis form and permit the user to design the desired type of synthetic speech for analysis. The major restriction on the character of the synthetic speech is caused by a limited choice of excitation waveforms. For both programs the excitation is binary, corresponding essentially to on-versus-off excitation.

The following input parameters may be selected for both programs:

- Synthetic vocal tract specification
  - Synthesizer size (number of taps)
  - Recursive coefficients
- Excitation function specification
  - Number of pitch pulses
  - Location of pitch pulses
- Analysis specification
  - Analysis block size
  - Analyzer size (number of LPC parameters).

Program CPMP5, in addition, permits vocal tract zeros to be present in the synthetic speech. Consequently, this program requires the number of zeros and the zero coefficients, i.e., the moving-average filter tap values. Program CPMP6 does not include zero modeling; however, it does model the case when the excitation cannot be completely stopped but remains at a fixed level. Consequently, with this program it is necessary

Preceding page blank

to specify the excitation off level (referenced to unity, which is the magnitude of the standard pitch pulse).

The following outputs are available from both of these programs:

- Synthetic speech
- Residual
- LPC parameter estimates
- Residual energy
- Normalized residual energy.

If so desired, the user can compare the estimated LPC parameters with the true synthesizer coefficients and observe the residual to look for the presence of the excitation pulses. The normalized residual energy is the test parameter that has been suggested for pitch extraction. This parameter should take on a very low value during force-free periods.

The listings for programs CPMP5 and CPMP6 are given in Figures A-1 and A-2, respectively.

```
>LOAD CPMP5
>LIST
001PRINT"THIS PROGRAM SYNTHESIZES INPUT,COMPUTES CORRELATION"
003PRINT"AND ERROR SIGNAL BY ATAL METHOD"
004PRINT"ENTER NUMBER OF DATA SAMPLES"
005INPUT N1
010PRINT"ENTER SYNTHESIZER SIZE "
011INPUT P1
012DIM T(1:P1)
013PRINT"ENTER COEFFICIENTS"
014FOR J=1 TO P1
015INPUT T(J)
016NEXT J
017PRINT"ENTER NUMBER OF ZEROS"
018INPUT N3
021PRINT"IF DESIRE ERROR AND SIGNAL PRINTOUT TYPE 1"
022INPUT N9
023PRINT"ENTER NUMBER OF PITCH PULSES"
024INPUT N5
025PRINT"ENTER LOCATION OF PITCH PULSES"
026DIM P(1:N5)
027FOR I=1 TO N5
028INPUT P(I)
029NEXT I
032PRINT"ENTER ANALYSIS BLOCK SIZE"
033INPUT N
035DIM S(-P1-N3+1:N1)
040FOR I=1 TO P1+N3
045S(I-P1-N3)=0
050NEXT I
051PRINT"IF DESIRE LPC PRINTOUT TYPE 1"
052INPUT N7
070FOR I=1 TO N1
096S(I)=0
097FOR J=1 TO P1
098S(I)=S(I)+S(I-J)*T(J)
099NEXT J
101FOR J=1 TO N5
102IF I=P(J) THEN 105 ELSE 103
103NEXT J
104GO TO 115
105S(I)=S(I)+1
115NEXT I
120DIM G(-P1:N1)
```

Reproduced from best available copy.

FIGURE A-1 LISTING OF PROGRAM CPMP5

```
127DIM M(1:N3)
128PRINT"ENTER ZERO TAPS"
129FOR J=1 TO N3
130INPUT M(J)
131NEXT J
134FOR K=1 TO N1+P1+1
135L=K-P1-1
136W(L)=0
137FOR J=1 TO N3
139W(L)=W(L)+S(L-J+1)*M(J)
140NEXT J
141NEXT K
195N2=N1/N
196PRINT"ENTER ANALYZER SIZE"
197INPUT P
200DIM R(0:P,0:P,1:N2)
205FOR L=1 TO N2
210FOR J=1 TO P+1
215FOR K=1 TO P+1
220R(J-1,K-1,L)=0
225FOR I=(L-1)*N+1 TO L*N
230R(J-1,K-1,L)=R(J-1,K-1,L)+G(I-J+1)*G(I-K+1)
235NEXT I
240NEXT K
245NEXT J
400DIM A(1:P,1:N2)
410DIM V(1:P,1:P,1:N2)
430V(1,1,L)=SQR(R(1,1,L))
440FOR J=2 TO P
450V(1,J,L)=R(1,J,L)/V(1,1,L)
460NEXT J
470FOR I=2 TO P
480X=0
490FOR K=1 TO I-1
500X=X+V(K,I,L)↑2
510NEXT K
520V(I,I,L)=SQR(R(I,I,L)-X)
530FOR J=1 TO P
540IF J<1 THEN 550 ELSE 570
550V(I,J,L)=0
560GO TO 630
570IF J=1 THEN 630 ELSE 580
580X=0
590FOR K=1 TO I-1
600X=X+V(K,I,L)*V(K,J,L)
610NEXT K
620V(I,J,L)=(R(I,J,L)-X)/V(I,I,L)
```

FIGURE A-1 LISTING OF PROGRAM CPMP5 (Continued)

```
630NEXT J
640NEXT I
650DIM W(1:P,1:N2)
660W(1,L)=R(0,1,L)/V(1,1,L)
670FOR I=2 TO P
680X=0
690FOR K=1 TO I-1
700X=X+V(K,I,L)*W(K,L)
702NEXT K
704W(I,L)=(R(0,I,L)-X)/V(I,I,L)
706NEXT I
708A(P,L)=W(P,L)/V(P,P,L)
710FOR I=1 TO P-1
712J=P-I
714X=0
716FOR K=J+1 TO P
718X=X+V(J,K,L)*A(K,L)
720NEXT K
722A(J,L)=(W(J,L)-X)/V(J,J,L)
724NEXT I
725IF N7=1 THEN 726 ELSE 740
726FOR J=1 TO P
728PRINT J, A(J,L)
730NEXT J
732PRINT"J","A"
740DIM E(1:N1)
750DIM F(1:N1)
751FOR I=(L-1)*N+1 TO L*N
752F(I)=0
760FOR J=1 TO P
770F(I)=F(I)+A(J,L)*G(I-J)
780NEXT J
790E(I)=G(I)-F(I)
800IF N9=1 THEN 810 ELSE 840
810PRINT I,E(I),G(I)
815NEXT I
820PRINT"I","ERROR","SIGNAL"
840PRINT R(0,1,L),R(0,2,L)
841PRINT"CORRELATION VECTOR"
842PRINT R(1,1,L),R(1,2,L)
843PRINT R(2,1,L),R(2,2,L)
844PRINT"CORRELATION MATRIX"
845Z1=0
846FOR J=1 TO P
847Z1=Z1+A(J,L)*R(0,J,L)
848NEXT J
849E=1-Z1/R(0,0,L)
850E1=R(0,0,L)-Z1
851PRINT E, E1
852PRINT"NORM ERR ENERGY","ERR ENERGY"
860 NEXT L
870END
```

FIGURE A-1 LISTING OF PROGRAM CPMP5 (Concluded)

Reproduced from best available copy.

```
>LOAD CPMP6
>LIST
001PRINT"THIS PROGRAM SYNTHESIZES INPUT,COMPUTES CORRELATION"
003PRINT"AND ERROR SIGNAL BY ATAL METHOD"
004PRINT"ENTER NUMBER OF DATA SAMPLES"
005INPUT N1
010PRINT"ENTER SYNTHESIZER SIZE "
011INPUT P1
012DIM T(1:P1)
013PRINT"ENTER COEFFICIENTS"
014FOR J=1 TO P1
015INPUT T(J)
016NEXT J
017PRINT"ENTER EXCITATION OFF LEVEL"
018INPUT N4
021PRINT"IF DESIRE ERROR AND SIGNAL PRINTOUT TYPE 1"
022INPUT N9
023PRINT"ENTER NUMBER OF PITCH PULSES"
024INPUT N5
025PRINT"ENTER LOCATION OF PITCH PULSES"
026DIM P(1:N5)
027FOR I=1 TO N5
028INPUT P(I)
029NEXT I
032PRINT"ENTER ANALYSIS BLOCK SIZE"
033INPUT N
035DIM S(-P1:N1)
040FOR I=1 TO P1+1
045S(I-P1-1)=0
050NEXT I
051PRINT"IF DESIRE LPC PRINTOUT TYPE 1"
052INPUT N7
070FOR I=1 TO N1
096S(I)=0
097FOR J=1 TO P1
098S(I)=S(I)+S(I-J)*T(J)
099NEXT J
100S(I)=S(I)+N4
101FOR J=1 TO N5
102IF I=P(J) THEN 105 ELSE 103
103NEXT J
104GO TO 115
105S(I)=S(I)+1
115NEXT I
195N2=N1/N
```

Reproduced from best available copy.

FIGURE A-2 LISTING OF PROGRAM CPMP6

```
196PRINT"ENTER ANALYZED SIZE"
197INPUT P
200DIM R(0:P,0:P,1:N2)
205FOR L=1 TO N2
210FOR J=1 TO P+1
215FOR K=1 TO P+1
220R(J-1,K-1,L)=0
225FOR I=(L-1)*N+1 TO L*N
230R(J-1,K-1,L)=R(J-1,K-1,L)+S(I-J+1)*S(I-K+1)
235NEXT I
240NEXT K
245NEXT J
400DIM A(1:P,1:N2)
410DIM V(1:P,1:P,1:N2)
430V(1,1,L)=SQR(R(1,1,L))
440FOR J=2 TO P
450V(1,J,L)=R(1,J,L)/V(1,1,L)
460NEXT J
470FOR I=2 TO P
480X=0
490FOR K=1 TO I-1
500X=X+V(K,I,L)↑2
510NEXT K
520V(I,I,L)=SQR(R(I,I,L)-X)
530FOR J=1 TO P
540IF J<1 THEN 550 ELSE 570
550V(I,J,L)=0
560GO TO 630
570IF J=1 THEN 630 ELSE 580
580X=0
590FOR K=1 TO I-1
600X=X+V(K,I,L)*V(K,J,L)
610NEXT K
620V(I,J,L)=(R(I,J,L)-X)/V(I,I,L)
630NEXT J
640NEXT I
650DIMW(1:P,1:N2)
660W(1,L)=R(0,1,L)/V(1,1,L)
670FOR I=2 TO P
680X=0
690FOR K=1 TO I-1
700X=X+V(K,I,L)*W(K,L)
702NEXT K
704W(I,L)=(R(0,I,L)-X)/V(I,I,L)
706NEXT I
```

FIGURE A-2 LISTING OF PROGRAM CPMP6 (Continued)

```
708A(P,L)=W(P,L)/V(P,P,L)
710FOR I=1 TO P-1
712J=P-I
714X=0
716FOR K=J+1 TO P
718X=X+V(J,K,L)*A(K,L)
720NEXT K
722A(J,L)=(W(J,L)-X)/V(J,J,L)
724NEXT I
725IF N7=1 THEN 726 ELSE 740
726FOR J=1 TO P
728PRINT J, A(J,L)
730NEXT J
732PRINT"J","A"
740DIM E(1:N1)
750DIM F(1:N1)
751FOR I=(L-1)*N+1 TO L*N
752F(I)=0
760FOR J=1 TO P
770F(I)=F(I)+A(J,L)*S(I-J)
780NEXT J
790E(I)=S(I)-F(I)
800IF N9=1 THEN 810 ELSE 840
810PRINT I,E(I),S(I)
815NEXT I
820PRINT"I","ERROR","SIGNAL"
840PRINT R(0,1,L),R(0,2,L)
841PRINT"CORRELATION VECTOR"
842PRINT R(1,1,L),R(1,2,L)
843PRINT R(2,1,L),R(2,2,L)
844PRINT"CORRELATION MATRIX"
845Z1=0
846FOR J=1 TO P
847Z1=Z1+A(J,L)*R(0,J,L)
848NEXT J
849E=1-Z1/R(0,0,L)
850E1=R(0,0,L)-Z1
851PRINT E, E1
852PRINT"NORM ERR ENERGY","ERR ENERGY"
860 NEXT L
870END
```

FIGURE A-2 LISTING OF PROGRAM CPMP6 (Concluded)

Reproduced from best available copy.

REFERENCES

1. J. L. Flanagan, Speech Analysis, Synthesis and Perception, Second Edition, p. 184 (Springer Verlag, New York, New York, 1972).

2. B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," Bell Sys. Tech. J., Vol. 49, No. 8, pp. 1973-1986 (October 1970).

3. J. J. Spilker, Jr., and D. T. Magill, "The Delay-Lock Discriminator--An Optimum Tracking Device," Proc. IRE, Vol. 49, pp. 1403-1416 (September 1961).

4. B. Gold and L. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acous. Soc. Am., Vol. 46, No. 2 (Part 2), pp. 442-448 (1969).

5. J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. Audio and Electroacous., Vol. AU-20, No. 5, pp. 367-377 (December 1972).

6. U. Mengali, "Acquisition Behavior of Generalized Tracking Systems in the Absence of Noise," IEEE Trans. Comm. Tech., Vol. COM-21, pp. 820-826 (July 1973).

7. W. C. Lindsey, Synchronization Systems in Communication and Control, p. 463 (Prentice-Hall, Englewood Cliffs, New Jersey, 1972).

8. J. D. Markel, A. H. Gray, Jr., and H. Waksita, "Linear Prediction of Speech-Theory and Practice," notes for a short course, Speech Communication Research Laboratory, Santa Barbara, California, 30 July to 1 August, 1973.

9. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acous. Soc. Am., Vol. 50, No. 2 (Part 2), pp. 637-655 (August 1971).

10. J. B. Allen and T. H. Curtis, "Automatic Extraction of Glottal Pulses by Linear Estimation," Conference Record, p. 36, Acoustic Society of America convention, Los Angeles, California, 29 October to 2 November, 1973.

11. R. Zetterberg, "Estimation of Parameters for a Linear Difference Equation Application to EEG Analysis," Mathematical Biosciences, Vol. 5, pp. 227-275 (American Elsevier Co., 1964).

12. A. N. Sobakin, "Digital Computer Determination of the Formant Parameters of the Vocal Tract from a Speech Signal," Soviet Physics--Acoustics, Vol. 18, No. 1, pp. 84-90 (July-September 1972).

13. C.G.M. Fant, "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies," For Roman Jakobson, M. Halle, H. Lunt, and H. MacLeun, eds., pp. 109-120 (The Hague; Mouton, 1956).

14. R. M. Lerner, "Band-Pass Filters with Linear Phase," Proc. IEEE, Vol. 52, pp. 249-268 (March 1964).

15. S. S. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-22, No. 2, pp. 135-141 (April 1974).

16. M. R. Schroeder and E. E. David, Jr., "A Vocoder for Transmitting 10 Kc/s Speech over a 3.5 Kb/s Channel," Acustica, Vol. 10, pp. 35-43 (1960).

17. R. A. McDonald, "Signal-to-Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems--Particular Applications to Voice Signals," Bell Sys. Tech. J., pp. 1123-1151 (September 1966).

18. J. Melsa et al., "Development of a Configuration Concept of a Speech Digitizer Based on Adaptive Estimation Techniques," Final Report, Southern Methodist University, Dallas, Texas (1973).

19. J. G. Dunn, "An Experimental 9600 Bits/s Voice Digitizer Employing Adaptive Prediction," IEEE Trans. Comm. Tech., Vol. COM-19, No. 6, pp. 1021-1032 (December 1971).

20. F. Itakura and S. Saito, "Analysis Synthesis Telephony Based upon the Maximum Likelihood Method," Reports Sixth International Congress on Acoustics, Tokyo, Japan (August 1968).

21. J. D. Markel, "Digital Inverse Filtering--A New Tool for Formant Trojectory Estimation," IEEE Trans. Audio and Electroacoust., Vol. AU-20, pp. 129-137 (June 1972).

22. J. I. Makhoul and J. J. Wolf, "Linear Prediction and the Spectral Analysis of Speech," BBN Report No. 2304, Bolt, Beranek and Newman, Inc., Cambridge, Massachusetts (August 1972).

23. N. Levinson, "The Wiener RMS Error Criterion in Filter Design and Prediction," J. Math. Phys., Vol. 25, No. 4, pp. 261-278 (1974). Also in N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series (MIT Press, Cambridge, Massachusetts, 1966).

24. E A. Robinson, Statistical Communication and Detection (Hafner Publishing Co., New York, New York, 1967).

25. J. D. Markel and A. H. Gray, Jr., "On Autocorrelation Equations as Applied to Speech Analysis," IEEE Trans. Audio and Electroacoust., Vol. AU-21, No. 2, pp. 69-79 (April 1973).

26. B. Gold and J. Tierney, "Digitized Voice-Excited Vocoder for Telephone-Quality Inputs, Using Bandpass Sampling of the Baseband Signal," J. Acous. Soc. Am., Vol. 37, pp. 753-754 (April 1965).

27. F. de Jager, "Deltamodulation: A Method of PCM Transmission Using a 1-Unit Code," Philips Res. Report 7, pp. 442-446 (1952).

28. A. Tomozawa and H. Kaneko, "Companded Delta Modulation for Telephone Transmission," IEEE Trans. Comm. Tech., Vol. COM-16, No. 1, pp.149-157 (February 1968).

29. S. J. Brolin and J. M. Brown, "Companded Delta Modulation for Telephony," IEEE Trans. Comm. Tech., Vol. COM-16, No. 1, pp. 157-162 (February 1968).

30. J. A. Greefkes, "Code Modulation System for Voice Signals Using Bit Rate Below 8 Kb/s," International Communication Conference Record, pp. 46.8-46.11 (1973).

31. M. R. Winkler, "High Information Delta Modulation," IEEE International Convention Record, Pt. 8, pp. 260-265 (1963).

32. J. E. Abate, "Linear and Adaptive Delta Modulation," Proc. IEEE, Vol. 55, No. 3, pp. 298-308 (March 1967).

33. N. S. Jayant, "Adaptive Delta Modulation with a One-Bit Memory," Bell Sys. Tech. J., Vol. 49, No. 3, pp. 321-342 (March 1970).

34. D. Middleton, An Introduction to Statistical Communication Theory, Chapter 5 (McGraw-Hill, New York, New York, 1960).

35. R. E. Bogner and M. Nashed, "Linear Harmonic Generation," Proc. IEEE, Vol. 120, No. 11, pp. 1328-1330 (November 1973).

36. F. Itakura and S. Saito, "Digital Filtering Techniques for Speech Analysis and Synthesis," Report Seventh International Congress on Acoustics, pp. 637-655, Budapest, Hungary (1971).

37. O. Fujimura, "An Approximation to Voice Aperiodicity," IEEE Trans. Audio and Electroacoust., Vol. AU-16, No. 1, pp. 68-72 (March 1968).